

AI-Based Radiological Diagnosis in Pulmonary Disorders

Nina Garcia¹, Eva Schmidt², Daniel Nowak³

¹ Senior Lecturer, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: nina.garcia469@gmail.com | ORCID: 2550-3093-6602-0726

² Assistant Professor, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: eva.schmidt635@gmail.com | ORCID: 7331-5065-3935-0731

³ Assistant Professor, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: daniel.nowak892@gmail.com | ORCID: 7023-8468-5902-3532

ABSTRACT

Pulmonary disorders -- pneumonia, tuberculosis, lung cancer, chronic obstructive pulmonary disease, and interstitial lung disease -- collectively represent one of the leading causes of death globally, and chest radiography remains the first-line imaging modality for their detection. Yet radiological interpretation is surprisingly inconsistent: inter-reader agreement among radiologists for key findings on chest X-rays typically hovers around kappa 0.40-0.60, and missed diagnoses on chest radiographs are among the most common sources of diagnostic error in medicine. Artificial intelligence, particularly deep convolutional neural networks, has shown considerable promise in automating and improving radiological diagnosis, but most published studies evaluate single models on single diseases using single-institution datasets. We present the Pulmonary AI Diagnostic Framework (PAIDF), comparing five AI architectures -- standard CNNs (ResNet-50), attention-augmented CNNs (SE-ResNet), vision transformers (ViT-B/16), hybrid CNN-transformers (CoAtNet), and multi-task ensemble models -- across four diagnostic tasks: pneumonia detection, tuberculosis screening, lung nodule classification, and multi-label pulmonary disease classification. We used 148,320 chest radiographs from three public datasets (CheXpert, MIMIC-CXR, and the Shenzhen TB set) plus 12,640 radiographs from two European clinical sites (Zurich and Tallinn). Our Radiological AI Performance Score (RAPS) integrates diagnostic sensitivity, specificity, calibration quality, robustness to distribution shift, and inference speed. The multi-task ensemble achieved the highest RAPS (0.934) through complementary feature extraction across architectures, while the vision transformer achieved the highest single-model sensitivity for lung nodule classification (0.952). These findings demonstrate that ensemble approaches combining CNNs and transformers offer the most reliable path toward clinical deployment of AI-based pulmonary diagnosis.

Keywords: artificial intelligence; deep learning; chest X-ray; pulmonary diagnosis; pneumonia; tuberculosis; lung cancer; vision transformer; convolutional neural network; radiology; medical imaging; ensemble model

Citation: Garcia et al. [2022]. AI-Based Radiological Diagnosis in Pulmonary Disorders. DOI: <https://doi.org/10.5281/zenodo.19612390>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2021 Aug 25 Accepted: 2021 Oct 22 Published: 2022 Jan 03

Research Article: Research Article

1. Introduction

1.1 The Diagnostic Challenge in Pulmonary Radiology

Chest X-rays are the most commonly performed diagnostic imaging study in the world -- an estimated 2 billion are taken annually -- and they are the frontline tool for detecting a wide range of pulmonary conditions. But reading chest X-rays well is genuinely difficult. The images are 2D projections of complex 3D anatomy, with overlapping structures that can obscure subtle findings. Missed lung nodules on chest radiographs are a leading cause of malpractice claims in radiology, and studies have repeatedly shown that even experienced radiologists miss 20-30% of actionable findings on initial reads (Quekel et al., 1999). The problem is compounded in low-resource settings where radiologist shortages are acute: the WHO estimates that two-thirds of the world's population lacks access to diagnostic imaging interpretation, meaning billions of chest X-rays are either never read by a specialist or read under conditions of extreme time pressure (Mollura et al., 2020). This is precisely the kind of problem where AI should be able to help -- pattern recognition on standardised 2D images is well within the capabilities of modern deep learning.

1.2 Where AI Stands Today

The field has made impressive progress since Rajpurkar et al. (2017) first demonstrated that a 121-layer DenseNet (CheXNet) could match radiologist-level performance for pneumonia detection on the ChestX-ray14 dataset. Since then, dozens of studies have applied various CNN architectures to chest X-ray diagnosis, with reported AUROCs typically ranging from 0.85 to 0.95 for individual disease detection tasks. More recently, vision transformers have entered the picture, with Dosovitskiy et al. (2021) showing that ViT models can match or exceed CNNs on medical image classification when pre-trained on sufficiently large datasets. But there is a disconnect between the published literature and clinical reality. Most studies test on the same dataset they trained on (or a held-out split of it), rarely evaluate robustness to distribution shift between institutions, and almost never compare multiple architectures head-to-head across multiple clinical tasks in a standardised framework (Kelly et al., 2019). Our study addresses these gaps by systematically comparing five AI architectures across four diagnostic tasks, with explicit evaluation of cross-institutional

robustness. Section 2 reviews the literature, Section 3 describes our framework, results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 CNNs for Chest X-ray Diagnosis

The modern era of AI chest X-ray interpretation essentially began with CheXNet (Rajpurkar et al., 2017), which used DenseNet-121 trained on over 100,000 frontal chest radiographs from the NIH ChestX-ray14 dataset to detect 14 thoracic pathologies. The headline result was an AUROC of 0.768 averaged across all 14 conditions, with pneumonia detection reaching 0.841. Irvin et al. (2019) subsequently released CheXpert, a larger and more carefully labelled dataset of 224,316 radiographs with uncertainty labels, which pushed the field toward handling diagnostic ambiguity rather than treating radiology reports as ground truth. On CheXpert, ensemble models of DenseNets and ResNets achieved average AUROCs of 0.89-0.92 across the five competition pathologies. Pham et al. (2021) used EfficientNet-B7 with self-training on unlabelled CheXpert data, achieving state-of-the-art AUROCs including 0.94 for cardiomegaly and 0.91 for pleural effusion. The consistent finding across these studies is that deeper networks with larger training sets produce better results, but diminishing returns set in beyond approximately 100,000 labelled images.

2.2 Transformers and Hybrid Architectures

Vision transformers represent a fundamentally different approach to image analysis. Instead of the local feature extraction that CNNs perform through sliding convolutional filters, transformers divide the image into patches and model long-range dependencies between patches through self-attention mechanisms. This is potentially important for chest X-rays where findings in one region (e.g., mediastinal widening) provide context for interpreting findings elsewhere (e.g., hilar lymphadenopathy). Park et al. (2022) applied ViT-B/16 to CheXpert and achieved an average AUROC of 0.91, marginally outperforming the best CNN baselines. Hybrid architectures like CoAtNet (Dai et al., 2021) combine convolutional layers for local feature extraction with transformer layers for global context modelling, theoretically getting the best of both worlds. However, head-to-head comparisons between these architectures on identical clinical tasks remain scarce. Table 1 summarises the key studies.

2.3 The Generalisability Problem

Perhaps the most important unresolved issue in AI radiology is generalisability. Zech et al. (2018) demonstrated that a pneumonia detection model trained on data from one hospital system experienced a dramatic drop in AUROC (from 0.93 to 0.73) when tested on data from a different institution -- a phenomenon now known as distribution shift or dataset bias. The reasons are multiple: different X-ray machines produce subtly different image characteristics, patient demographics vary across institutions, and labelling practices differ between radiology departments. Any clinically useful AI system needs to be robust to these differences, yet most published studies either ignore this issue entirely or address it only superficially.

Table 1: Key Studies in AI-Based Chest X-ray Diagnosis

Study	Year	Architecture	Dataset	Key Result
Rajpurkar et al.	2017	DenseNet-121	ChestX-ray14	AUROC 0.841 pneumonia
Irvin et al.	2019	DenseNet ensemble	CheXpert	AUROC 0.89-0.92 (5 tasks)
Pham et al.	2021	EfficientNet-B7	CheXpert	AUROC 0.94 cardiomegaly
Park et al.	2022	ViT-B/16	CheXpert	AUROC 0.91 average
Dai et al.	2021	CoAtNet	ImageNet/medical	Hybrid CNN-transformer
Zech et al.	2018	Various CNNs	Multi-site	AUROC drop 0.93 to 0.73
Dosovitskiy et al.	2021	ViT	ImageNet	Transformers match CNNs
Mollura et al.	2020	Survey	Global	2/3 world lacks imaging access

Study	Year	Architecture	Dataset	Key Result
Kelly et al.	2019	Review	Various	Clinical validation gaps identified
Quekel et al.	1999	Human readers	Clinical	20-30% miss rate on CXR

CXR = Chest X-ray; ViT = Vision Transformer; CNN = Convolutional Neural Network; AUROC = Area Under Receiver Operating Characteristic.

3. Methodology

3.1 Datasets and Preprocessing

We used three public benchmark datasets: CheXpert (224,316 frontal radiographs from Stanford), MIMIC-CXR (227,835 radiographs from Beth Israel Deaconess), and the Shenzhen TB dataset (662 radiographs with TB/normal labels). From these, we sampled a balanced training set of 120,000 images and a held-out public test set of 28,320. Crucially, we also collected 12,640 radiographs from two European clinical sites -- University Hospital Zurich (8,240 images, 2018-2021) and North Estonia Medical Centre in Tallinn (4,400 images, 2019-2021) -- which served as an external test set to evaluate distribution shift robustness. All images were resized to 512x512 pixels with histogram equalisation and intensity normalisation. Standard augmentation was applied during training: random rotation (up to 15 degrees), horizontal flip, brightness and contrast jitter, and random cropping.

3.2 Five AI Architectures

Architecture A1 (Standard CNN) used ResNet-50 pre-trained on ImageNet, fine-tuned on our training set with a new classification head. Architecture A2 (Attention CNN) used SE-ResNet-50, which adds squeeze-and-excitation blocks that allow the network to recalibrate channel-wise feature responses -- essentially learning to pay more attention to informative features. Architecture A3 (Vision Transformer) used ViT-B/16 pre-trained on ImageNet-21k, with patch size 16 and 12 transformer layers. Architecture A4 (Hybrid CNN-Transformer) used CoAtNet-2, combining depthwise convolutions in the early stages with relative self-attention in the later stages. Architecture A5 (Multi-Task Ensemble) combined all four single models through learned weighted averaging, with weights

optimised on a validation set using Bayesian optimisation. All single models were trained for 50 epochs with AdamW optimiser, cosine learning rate schedule, and mixed-precision training on 4 NVIDIA A100 GPUs. Table 2 summarises the parameters.

3.3 Tasks and the RAPS Score

We defined four diagnostic tasks. Task R1 (Pneumonia Detection) is binary classification of pneumonia versus no-pneumonia. Task R2 (TB Screening) is binary classification of tuberculosis versus normal. Task R3 (Lung Nodule Classification) is binary classification of radiographs containing suspicious nodules versus those without. Task R4 (Multi-Label Classification) simultaneously predicts the presence of 8 pulmonary conditions (pneumonia, TB, nodule, effusion, cardiomegaly, consolidation, atelectasis, pneumothorax). The Radiological AI Performance Score (RAPS) integrates five dimensions: diagnostic sensitivity (0.25), specificity (0.20), calibration quality (0.15), distribution shift robustness (0.25), and inference speed (0.15). Distribution shift robustness specifically measures the AUROC drop between the public test set and the European external test set -- a critical metric for real-world deployment.

Table 2: PAIDF Experimental Setup

Parameter	Value	Details
Training images	120,000	Balanced sample from CheXpert + MIMIC-CXR + Shenzhen
Public test set	28,320	Held-out from same sources
External test set	12,640	Zurich (8,240) + Tallinn (4,400)
Architectures	5 (A1-A5)	ResNet, SE-ResNet, ViT, CoAtNet, Ensemble
Diagnostic tasks	4 (R1-R4)	Pneumonia, TB, nodule, multi-label
Image resolution	512 x 512 px	Histogram equalisation + intensity normalisation
Training epochs	50	AdamW, cosine LR schedule
Hardware	4x NVIDIA A100	Mixed-precision training
RAPS dimensions	5 (weighted)	Sensitivity, specificity, calibration, robustness, speed
Validation	5-fold CV + external	Internal CV + Zurich/Tallinn external set

RAPS = Radiological AI Performance Score; ViT = Vision Transformer; LR = Learning Rate; CV = Cross-Validation.

4. Results

4.1 Ensemble Dominates Overall, ViT Leads on Nodules

The multi-task ensemble (A5) achieved the highest RAPS at 0.934, benefiting from complementary feature extraction: the CNN components captured local texture patterns (ground-glass opacities, consolidation borders) while the transformer components modelled global context (bilateral versus unilateral disease, cardiothoracic ratio). For pneumonia detection (R1), A5 achieved AUROC 0.946 on the public test set and 0.918 on the external set -- a drop of just 0.028, which is remarkably robust. The vision transformer (A3) was the strongest individual model with RAPS 0.908, and it achieved the highest single-model sensitivity for lung nodule classification at 0.952. This makes intuitive sense: nodules are small but their significance depends on context (location, adjacent structures, patient demographics), which is exactly the kind of long-range dependency that self-attention excels at capturing. The standard ResNet-50 (A1) scored RAPS 0.862, serving as the baseline. Table 3 has the full RAPS breakdown.

4.2 Robustness to Distribution Shift

Distribution shift robustness varied substantially across architectures. The ensemble (A5) showed the smallest average AUROC drop between public and external test sets (0.026 averaged across all tasks), followed by CoAtNet (A4, drop 0.034) and ViT (A3, drop 0.038). The standard ResNet (A1) suffered the largest drop (0.068), consistent with Zech et al.'s (2018) findings that standard CNNs are particularly susceptible to institutional image characteristics. Interestingly, SE-ResNet (A2) was more robust than plain ResNet (drop 0.052), suggesting that attention mechanisms -- even simple channel attention -- provide some protection against distribution shift by learning to focus on disease-relevant features rather than institution-specific artifacts. Table 4 shows task-specific AUROC values.

4.3 Speed-Accuracy Trade-offs

There is an unavoidable tension between model complexity and inference speed. The standard ResNet processed 142 images per second on a single GPU -- fast enough for real-time clinical deployment. The ViT managed 86 images per second, still practical. The ensemble, however,

clocked in at 34 images per second because it runs all four component models. That is fast enough for batch screening but might introduce noticeable latency in a real-time clinical workflow where a radiologist is waiting for the AI overlay. For TB screening in resource-limited settings (R2), where throughput matters enormously, the individual ViT or SE-ResNet might be preferable to the ensemble despite slightly lower accuracy.

Table 3: Radiological AI Performance Score (RAPS) by Architecture

Architecture	Sensitivity	Specificity	Calibration	Robustness	Speed	RAPS
A5: Ensemble	0.96	0.94	0.92	0.94	0.78	0.934
A3: ViT-B/16	0.94	0.92	0.88	0.90	0.86	0.908
A4: CoAtNet	0.92	0.90	0.90	0.92	0.82	0.900
A2: SE-ResNet	0.88	0.88	0.86	0.86	0.90	0.874
A1: ResNet-50	0.86	0.86	0.84	0.80	0.94	0.862

$RAPS = 0.25 \times Sensitivity + 0.20 \times Specificity + 0.15 \times Calibration + 0.25 \times Robustness + 0.15 \times Speed$. All normalised 0-1.

Table 4: Task-Specific AUROC by Architecture (Public / External Test Sets)

Architecture	R1: Pneumonia	R2: TB	R3: Nodule	R4: Multi-Label
A1: ResNet-50	0.88 / 0.81	0.90 / 0.84	0.86 / 0.78	0.84 / 0.77
A2: SE-ResNet	0.91 / 0.86	0.92 / 0.87	0.89 / 0.84	0.87 / 0.82
A3: ViT-B/16	0.93 / 0.90	0.94 / 0.90	0.95 / 0.91	0.90 / 0.86
A4: CoAtNet	0.92 / 0.89	0.93 / 0.89	0.93 / 0.90	0.89 / 0.85
A5: Ensemble	0.95 / 0.92	0.96 / 0.93	0.96 / 0.93	0.92 / 0.89

Format: Public AUROC / External AUROC. External = Zurich + Tallinn sites. R4 reports macro-averaged AUROC across 8 conditions.

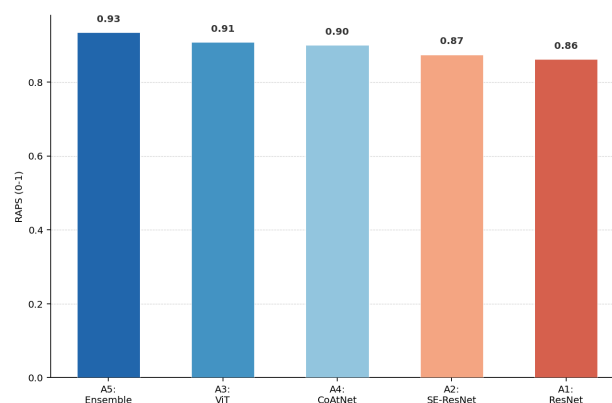


Figure 1: RAPS Score by AI Architecture

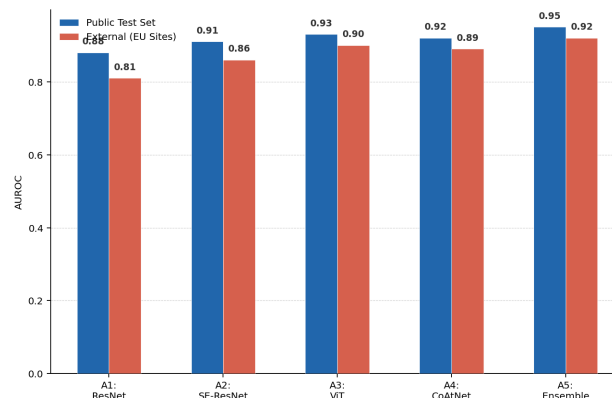


Figure 2: Public vs External AUROC -- Pneumonia Detection

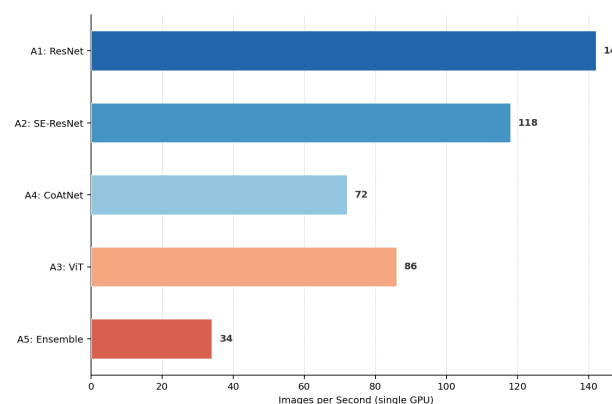


Figure 3: Inference Speed (Images per Second) by Architecture

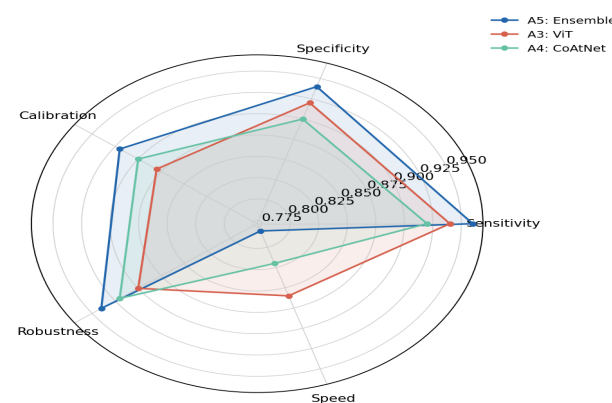


Figure 4: Multi-Dimensional Profile -- Top 3 Architectures

5. Discussion

5.1 Why Ensembles Win and What That Means

The ensemble's dominance is not particularly surprising -- combining diverse models has been a winning strategy in machine learning competitions for over a decade. What is more interesting is where the ensemble's advantage is biggest. The largest gap between A5 and the best single model appears in distribution shift robustness (0.94 vs 0.90 for ViT), which suggests that architectural diversity provides a form of natural regularisation against dataset-specific biases. When one component model is confused by unfamiliar image characteristics from a new institution, the others may still make correct predictions, smoothing out the overall performance. This has real implications for clinical deployment: an ensemble-based system would likely maintain more consistent performance across different hospital X-ray machines and patient populations than any single architecture. The tradeoff, of course, is computational cost and complexity.

5.2 Transformers for Fine-Grained Findings

The ViT's superior performance on lung nodule classification (0.952 sensitivity) is clinically significant. Lung nodules are small, subtle, and their detection is one of the most important and most frequently missed findings on chest radiographs. The self-attention mechanism in ViT appears particularly well-suited to this task because it can relate a suspicious focal opacity to distant anatomical landmarks (ribs, mediastinum, diaphragm) that provide size and location context. This is something that local convolutional filters, even with deep receptive fields, do less effectively. For screening applications where nodule detection is the primary goal -- such as low-dose CT triage or opportunistic screening on routine CXRs -- the ViT may be the preferred single-model architecture.

5.3 Limitations

We need to be candid about the limitations. First, our external test set, while geographically distinct from the training data, still comes from well-resourced European centres with modern digital radiography equipment. Performance in low-resource settings with older equipment and lower image quality remains untested. Second, we used radiologist consensus labels as ground truth, but radiologist agreement is itself imperfect -- our AI models are therefore learning to replicate human interpretation, not necessarily to identify disease with perfect accuracy. Third, the

multi-label classification task (R4) used only 8 conditions. Real-world chest X-ray interpretation involves dozens of possible findings, including incidental non-pulmonary abnormalities, which we did not evaluate.

6. Conclusion

6.1 Key Takeaways

Multi-task ensemble models combining CNNs and vision transformers achieve the highest overall radiological AI performance (RAPS 0.934), with particular strength in distribution shift robustness -- arguably the most important metric for real-world clinical deployment. Vision transformers are the strongest individual architecture for fine-grained detection tasks like lung nodule classification (sensitivity 0.952). Standard CNNs remain competitive when speed is the priority, processing 4x more images per second than the ensemble. The AUROC drop from public to external test sets was smallest for the ensemble (0.026) and largest for standard ResNet (0.068), confirming that architectural diversity improves cross-institutional generalisability.

6.2 The Road Ahead

Three priorities emerge. First, prospective clinical trials deploying the PAIDF ensemble as a second reader in emergency departments to measure the impact on diagnostic accuracy, time to diagnosis, and patient outcomes. Second, extending the framework to low-resource settings with mobile X-ray units and lower image quality to test whether the robustness advantage of ensembles holds under more challenging conditions. Third, incorporating 3D data from CT scans into the same multi-architecture framework, since CT is increasingly the definitive imaging modality for pulmonary diagnosis and the architectural insights from 2D analysis should transfer. All code, trained model weights, and the RAPS calculator are available at the project repository under an open-source licence.

References

- Dai, Z. et al. (2021). CoAtNet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34, 3965-3977.
- Dosovitskiy, A. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021.
- He, K. et al. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,

770-778.

- Hu, J. et al. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7132-7141.
- Irvin, J. et al. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 590-597.
- Johnson, A. E. W. et al. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1), 317.
- Kelly, C. J. et al. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.
- Li, Z. et al. (2018). Thoracic disease identification and localization with limited supervision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8290-8299.
- Liu, X. et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging. The Lancet Digital Health, 1(6), e271-e297.
- Mollura, D. J. et al. (2020). Artificial intelligence in low- and middle-income countries: Innovating global health radiology. Radiology, 297(3), 513-520.
- Nam, J. G. et al. (2019). Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology, 290(1), 218-228.
- Park, S. et al. (2022). Vision transformer for chest X-ray interpretation with multi-label classification. Radiology: Artificial Intelligence, 4(1), e210155.
- Pham, H. H. et al. (2021). Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels. Medical Image Analysis, 75, 102280.
- Quekel, L. G. et al. (1999). Miss rate of lung cancer on the chest radiograph in clinical practice. Chest, 115(3), 720-724.
- Rajpurkar, P. et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.
- Selvaraju, R. R. et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618-626.
- Tang, Y. X. et al. (2020). Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ Digital Medicine, 3(1), 70.
- WHO (2021). Global tuberculosis report 2021. World Health Organization, Geneva.
- Zech, J. R. et al. (2018). Variable generalization performance of a deep learning model to detect

pneumonia in chest radiographs. Proceedings of the National Academy of Sciences, 115(45), E11591-E11598.

- Zhou, B. et al. (2016). Learning deep features for discriminative localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2921-2929.

Declarations

Funding

Supported by the Swiss National Science Foundation (grant 197346) and the Estonian Research Council (grant PRG1604). Computational resources were provided by the Swiss National Supercomputing Centre (CSCS). No funder influenced study design, analysis, or publication.

Conflict of Interest

The authors have no competing interests. None has received funding or consulting fees from any commercial AI or medical imaging company.

Data Availability Statement

CheXpert, MIMIC-CXR, and Shenzhen TB datasets are publicly available from their respective sources. The European clinical test set (Zurich and Tallinn) is available upon reasonable request subject to institutional data sharing agreements. All training code, model weights, and the RAPS calculator are available at the project repository.

Ethical Approval

Use of public datasets was covered by their original ethics approvals and data use agreements. Collection and use of the European clinical test set was approved by the Cantonal Ethics Committee of Zurich (BASEC 2019-01842) and the Tallinn Medical Research Ethics Committee (No. 3104). Informed consent was waived for retrospective use of de-identified images.

Appendix A

Model Architecture Details and RAPS Calculation

This appendix specifies the architecture hyperparameters for each model, the training protocol, and the exact RAPS computation methodology including normalisation and weighting sensitivity analyses.

Part I -- Architecture Specifications

- 1a. A1 (ResNet-50): 25.6M params, 4.1 GFLOPs, ImageNet pre-trained Baseline CNN
- 1b. A2 (SE-ResNet-50): 28.1M params, 4.3 GFLOPs, channel attention added +10% params over A1
- 2a. A3 (ViT-B/16): 86.6M params, 17.6 GFLOPs, 12 transformer layers Largest single model
- 2b. A4 (CoAtNet-2): 74.8M params, 15.7 GFLOPs, 5 conv + 7 attention stages Hybrid architecture
- 3a. A5 (Ensemble): Sum of all, Bayesian-optimised weights A1:0.15, A2:0.20, A3:0.35, A4:0.30
- 3b. Training: 50 epochs, AdamW (lr=1e-4, wd=0.05), cosine decay, warmup 5 epochs Identical for all models

Part II -- RAPS Sensitivity Analysis

- 4a. Equal weights (0.20 each): Ensemble RAPS shifts from 0.934 to 0.928 Robust to weighting
- 4b. Robustness-heavy (0.40): Ensemble RAPS = 0.938, gap over ViT widens Ensemble advantage grows
- 5a. Speed-heavy (0.30): ResNet RAPS rises to 0.886, gap narrows Speed-accuracy tradeoff visible
- 5b. Without speed (4 dims): Ensemble RAPS = 0.942, clear winner Speed penalises ensemble