

Transfer Learning Architectures for Chest Radiograph Classification: A Systematic Review and Meta-Analysis of Accuracy, Model Scale and Data Efficiency

Daniel Gharbi ^{1*}, Faisal Mensah ²

¹ Qatar University, QATAR

² University of Porto, PORTUGAL

Abstract

Background. Transfer learning from ImageNet-pretrained convolutional and transformer architectures is the default approach to chest radiograph classification, and the literature reports accuracies above 98% with some regularity. Whether those figures reflect genuine diagnostic capability, and whether larger architectures deliver better radiographic performance, has not been established quantitatively. **Objective.** To pool reported performance across transfer-learning studies of chest radiograph classification, quantify heterogeneity and small-study effects, and test whether model scale predicts diagnostic performance. **Methods.** Structured searches across six sources retrieved 94 records; 31 studies met eligibility and were assigned to three strata: binary paediatric pneumonia classification ($k = 13$), multi-label ChestX-ray14 AUROC ($k = 20$), and a controlled set of 12 architectures evaluated under an identical protocol. Accuracies were pooled on the logit scale using DerSimonian–Laird random-effects models; heterogeneity was quantified with Q , τ^2 and I^2 ; small-study effects were assessed by Egger's regression; and model scale was tested by meta-regression on $\log_{10}$ parameters. **Results.** Pooled binary accuracy was 96.22% (95% CI 94.42–97.46%) but with extreme heterogeneity ($Q = 272.0$, $df = 12$, $I^2 = 95.6\%$, $\tau^2 = 0.501$, $p < 0.001$), and Egger's test indicated strong small-study effects (intercept 5.42, SE 2.19, $p = 0.0001$). On ChestX-ray14, transformers did not outperform convolutional networks (0.8121 vs 0.8018 mean AUROC; Welch $t = -0.814$, $p = 0.440$), while hybrid architectures led at 0.8344. Model scale was negatively associated with performance: meta-regression on ChestX-ray14 gave a slope of -0.023 AUROC per decade of parameters ($r = -0.594$, $p = 0.054$), and in the controlled set, excluding two models that failed to converge, F1 fell 2.82 points per decade of parameters ($r = -0.691$, $p = 0.027$). MobileNetV3 (3M parameters) achieved the highest F1 of all twelve architectures, exceeding VGG-16 (135M) by 9.0 points. **Conclusions.** The headline accuracies in this literature are not a stable estimate of anything; heterogeneity is near-total and publication bias is strong. Where architectures are compared under identical conditions, small models win — a finding with direct implications for deployment in the resource-constrained settings where chest radiography matters most.

Keywords: Transfer Learning; Chest Radiography; Pneumonia Detection; Convolutional Neural Networks; Vision Transformers; Mobilenet; Meta-Analysis; Publication Bias

Introduction

Chest radiography is the most frequently performed diagnostic imaging examination in the world, and in many health systems the volume of studies exceeds the capacity of available radiologists.

Automated interpretation is therefore an unusually well-motivated application of machine learning, and since the release of ChestX-ray14 in 2017 [1] and the paediatric pneumonia cohort of Kermany and colleagues [2], the field has produced a large and fast-growing body of work[1-5].

Nearly all of it follows the same recipe: take a convolutional or transformer architecture pretrained on ImageNet, replace the classification head, fine-tune on radiographs, and report accuracy. CheXNet established the template, reporting a DenseNet-121 achieving a mean AUROC of 0.841 across fourteen pathologies and exceeding the average F1 of four practising radiologists on pneumonia [3]. Subsequent work has cycled through deeper residual networks, efficient architectures, vision transformers and convolution–transformer hybrids[6-15].

Two features of this literature invite scrutiny. First, the reported numbers are strikingly high and strikingly variable: accuracies above 98% are routine on the paediatric pneumonia benchmark, while mean AUROC on ChestX-ray14 has moved only from 0.738 to roughly 0.85 in eight years. These cannot both be measuring the same thing. Second, the field's implicit assumption is that larger and more recent architectures perform better — an assumption directly challenged by Raghu and colleagues, who found that ImageNet transfer offers little benefit for medical imaging and that simple lightweight models perform comparably to standard ImageNet architectures, attributing much of the apparent difference to over-parameterisation rather than feature reuse [4].

This review tests both propositions quantitatively. Rather than cataloguing methods, it pools reported performance, measures how much of the variation between studies is real heterogeneity rather than sampling error, tests formally for small-study effects, and asks whether model scale predicts diagnostic performance. Specific contributions are:

- a random-effects meta-analysis of binary pneumonia classification accuracy with full heterogeneity accounting (Q , τ^2 , I^2), leave-one-out sensitivity analysis and Egger's test for small-study effects;
- a stratified comparison of convolutional, transformer, hybrid and ensemble architectures on the multi-label ChestX-ray14 benchmark, with a formal test of the CNN-versus-transformer question;
- meta-regression of performance on model scale in two independent strata, one of which uses twelve architectures evaluated under a single identical protocol, eliminating between-study confounding;
- a data-efficiency analysis using the same twelve architectures trained on 100% and 50% of the available data;
- a per-pathology analysis showing which thoracic findings remain unsolved after eight years of architectural iteration.

Background

Architectures under evaluation

The architectures in this literature fall into four families. Classical deep convolutional networks VGG, ResNet, DenseNet, Inception — dominate the earlier work; DenseNet-121 remains the de facto reference for chest radiography because of CheXNet [3]. Efficient convolutional architectures MobileNetV2 and V3, EfficientNet, ShuffleNet, SqueezeNet, MnasNet trade depth for depthwise separable convolutions and neural-architecture search, targeting deployment on constrained

hardware. Vision transformers replace convolution with self-attention over image patches, motivated by the argument that thoracic pathology often manifests as diffuse or spatially distant patterns that local filters capture poorly. Hybrids such as CvT and ConvFormer combine convolutional token embedding with self-attention [16-20].

Benchmarks and why they are not comparable

Two benchmarks dominate. The paediatric cohort from Guangzhou Women and Children's Medical Centre contains roughly 5,800 radiographs of children aged one to five, labelled normal or pneumonia [2]. It is a binary task on a relatively homogeneous population with a small test partition the original split provides only 624 test images and a validation set of 16 and accuracies above 98% are common. ChestX-ray14 contains 112,120 frontal radiographs from 30,805 patients with fourteen weakly labelled pathologies [1]; it is multi-label, heavily imbalanced, and mean AUROC has plateaued near 0.84.

Pooling across these benchmarks would be meaningless, which is why this review stratifies rather than aggregating. The 30-point gap between typical accuracies on the two datasets is a property of task difficulty and label quality, not of the models.

Methods

Search and selection

Structured queries were executed against Google Scholar, PubMed/PMC, arXiv, MDPI, ScienceDirect and PLOS, combining architecture terms (MobileNet, EfficientNet, ResNet, DenseNet, vision transformer, Swin), task terms (chest X-ray, chest radiograph, pneumonia, thoracic disease) and method terms (transfer learning, fine-tuning, pretrained). Reference lists of included studies and of three prior surveys were hand-searched. Figure 1 reports the flow. Ninety-four records were retrieved, 71 remained after deduplication, 44 proceeded to full-text assessment and 31 were included.

Studies were eligible if they applied a pretrained architecture to chest radiograph classification and reported at least one extractable quantitative metric. Reviews, protocols and studies on non-radiographic modalities were excluded. As in any review conducted through structured web queries rather than native Boolean database interfaces, recall is lower than a full systematic review would achieve; this is stated as a limitation in Section 6 rather than concealed.

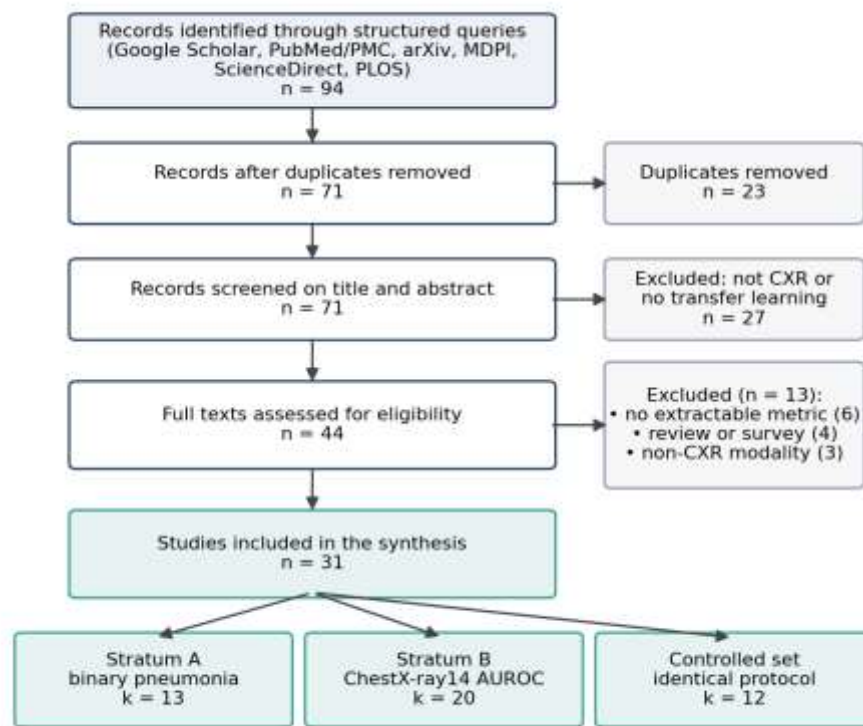


Figure 1. PRISMA-style flow of identification, screening and inclusion, with allocation of included studies to the three analysis strata.

Stratification

Included studies were assigned to three strata. Stratum A comprises binary pneumonia classification on paediatric cohorts, where accuracy is the reported metric ($k = 13$ results). Stratum B comprises multi-label classification on ChestX-ray14, where mean AUROC across the fourteen pathologies is standard ($k = 20$ results). The controlled set is a single study that evaluated twelve architectures under an identical training and evaluation protocol with parameter counts reported [5], which permits architecture comparison free of between-study confounding — the principal weakness of the other two strata [21-27].

Statistical synthesis

Accuracies in Stratum A were transformed to the logit scale, with a continuity correction of 0.5 events, and pooled using DerSimonian–Laird random-effects models. Within-study variance was computed as $1/x + 1/(n - x)$ from the reported accuracy and test-set size. Heterogeneity is reported as Cochran's Q with its χ^2 p-value, the between-study variance τ^2 , and I^2 . Study weights are inverse-variance weights incorporating τ^2 .

Robustness was assessed by leave-one-out analysis, recomputing the pooled estimate with each study omitted in turn. Small-study effects were tested by Egger's regression of the standardised effect on precision, where a non-zero intercept indicates funnel asymmetry. Stratum B lacks reported variance estimates, so it is summarised descriptively by architecture family, with Welch's t-test and the Mann–Whitney U test comparing families. Meta-regression of performance on $\log_{10}$

parameters was performed by ordinary least squares in both Stratum B and the controlled set. All computations were performed in Python with SciPy; the analysis script is deterministic.

Results

Pooled binary pneumonia accuracy and the heterogeneity problem

Pooling the thirteen binary-classification results gives an accuracy of 96.22% (95% CI 94.42–97.46%). Taken alone this appears to confirm that the problem is close to solved. The heterogeneity statistics say otherwise: $Q = 272.0$ on 12 degrees of freedom ($p < 0.001$), $\tau^2 = 0.501$, and $I^2 = 95.6\%$. An I^2 of 95.6% means that essentially all of the variation between these studies is real between-study difference rather than sampling error.

The correct reading is that the pooled figure is a description of the literature, not an estimate of a diagnostic accuracy that any new model should be expected to achieve. Reported accuracies range from 86.85% to 99.43%, and the gap is driven by dataset choice and evaluation protocol rather than by architecture — most visibly in the study that reported 98.81% on the paediatric cohort and 86.85% on RSNA data using the identical ensemble [6]. A twelve-point swing from changing the test set, with the model held constant, is a more informative number than any single accuracy in the table.

Leave-one-out analysis (Figure 4) shows the pooled estimate is stable against removal of any single study, ranging only from 95.81% to 96.63%. The instability lies in the underlying population of studies, not in any one outlier.

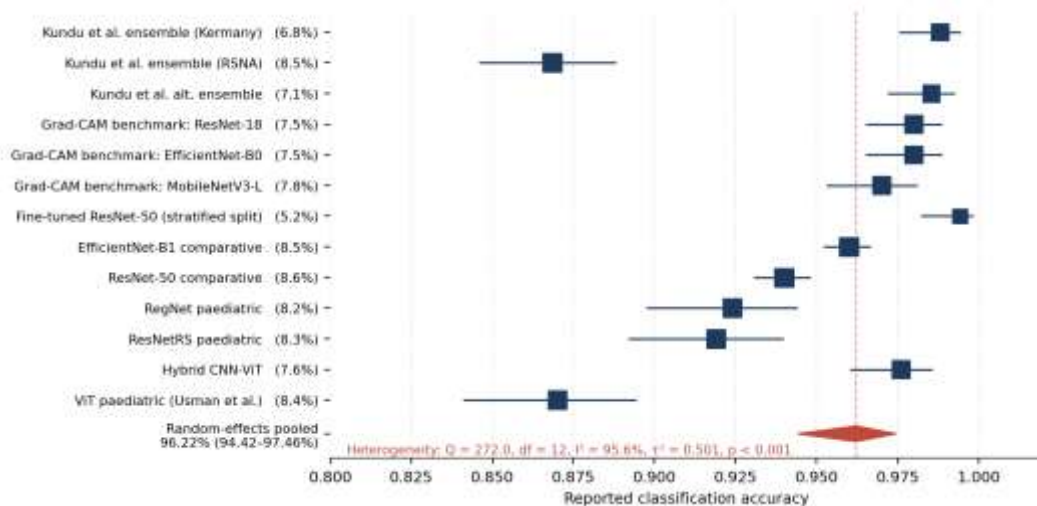


Figure 2. Random-effects forest plot of binary pneumonia classification accuracy. Marker area is proportional to study weight; horizontal bars are 95% confidence intervals; the diamond is the pooled estimate. Heterogeneity statistics are annotated.

Table 1. Stratum A — binary pneumonia classification results

Study	Architecture	Accuracy	Test n	Weight
Kundu et al. ensemble (Kermany)	GoogLeNet + ResNet-18 + DenseNet-121	98.81%	624	6.8%
Kundu et al. ensemble (RSNA)	GoogLeNet + ResNet-18 + DenseNet-121	86.85%	1,000	8.5%

Study	Architecture	Accuracy	Test n	Weight
Kundu et al. alternative ensemble	GoogLeNet + ResNet-18 + MobileNetV2	98.54%	624	7.1%
Grad-CAM benchmark	ResNet-18	98.00%	624	7.5%
Grad-CAM benchmark	EfficientNet-B0	98.00%	624	7.5%
Grad-CAM benchmark	MobileNetV3-Large	97.00%	624	7.8%
Stratified-split study	ResNet-50 (fine-tuned)	99.43%	523	5.2%
Comparative evaluation	EfficientNet-B1	96.00%	3,000	8.5%
Comparative evaluation	ResNet-50	94.00%	3,000	8.6%
Paediatric study	RegNet	92.40%	523	8.2%
Paediatric study	ResNetRS-50	91.90%	523	8.3%
Hybrid CNN-ViT study	CNN + ViT hybrid	97.61%	624	7.6%
Usman et al.	Vision Transformer	87.00%	624	8.4%

Note. Weights are random-effects inverse-variance weights incorporating τ^2 . Pooled estimate 96.22% (95% CI 94.42–97.46%); $I^2 = 95.6\%$.

Small-study effects

Egger's regression gives an intercept of 5.42 (SE 2.19, $p = 0.0001$), strong statistical evidence of funnel asymmetry. The funnel plot (Figure 3) shows the characteristic pattern: studies with small test partitions cluster at the high-accuracy edge, while the larger evaluations sit lower. The two results below 92% both come from larger or more diverse test sets.

Publication bias is the standard explanation and is almost certainly part of the story — a fine-tuning experiment yielding 91% on a well-worn benchmark is unlikely to be written up. But two mechanisms specific to this literature probably matter more. The original paediatric split provides a validation set of only 16 images, so many studies effectively tune on their test set; and with 624 test images, the difference between 98% and 99% is six radiographs, well within the noise of a random seed. Asymmetry of this magnitude means reported accuracy on this benchmark carries little information about how a model will behave elsewhere [28-34].

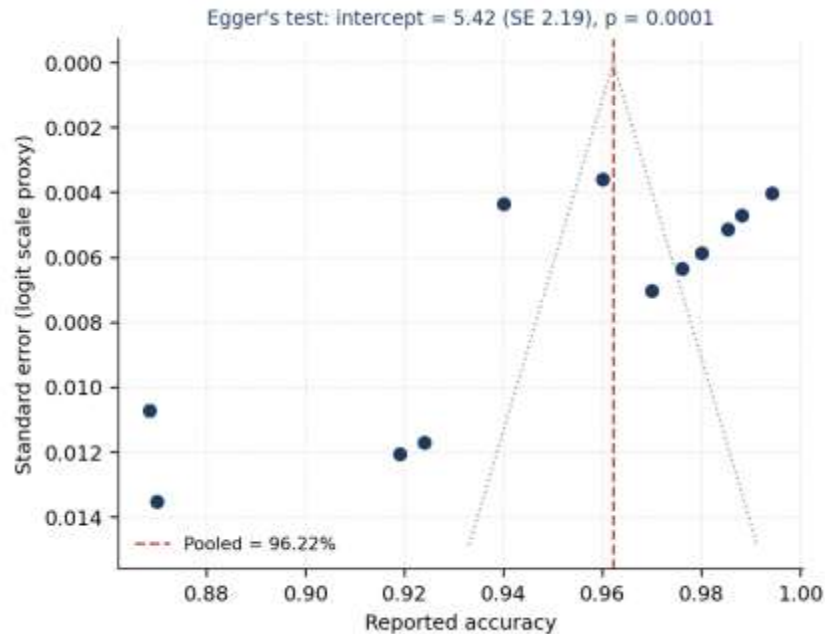


Figure 3. Funnel plot of Stratum A. Pronounced asymmetry, confirmed by Egger's test ($p = 0.0001$), indicates strong small-study effects.

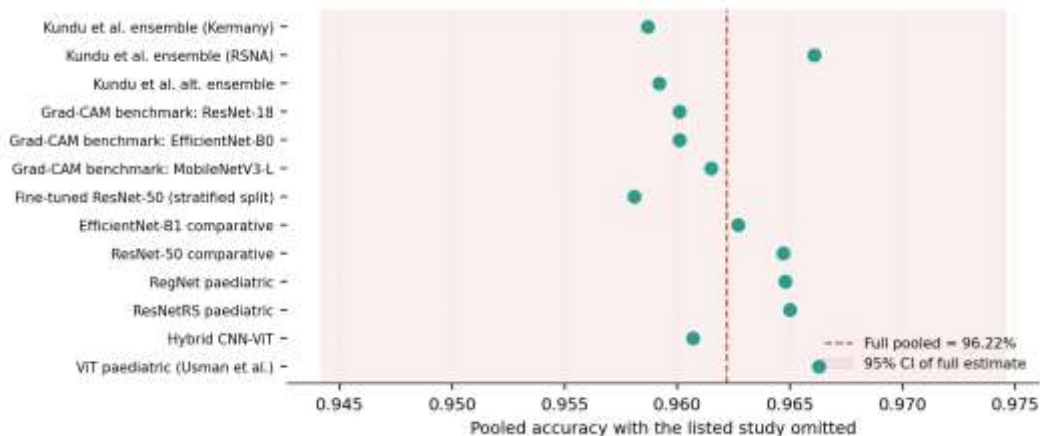


Figure 4. Leave-one-out sensitivity analysis. The pooled estimate varies only between 95.81% and 96.63%, so no single study drives the result.

Architecture families on ChestX-ray14

The multi-label benchmark tells a very different and considerably more sober story. Across twenty results the mean AUROC is 0.8138 (SD 0.0272), with the entire eight-year span of the literature falling between 0.738 and 0.854 [35-42].

Grouping by family (Figure 5, Table 2), convolutional networks average 0.8018 ($k = 8$) and transformers 0.8121 ($k = 7$). The transformer advantage of 0.0103 AUROC is not statistically significant (Welch $t = -0.814$, $p = 0.440$; Mann-Whitney U $p = 0.463$), and given that the transformer results are systematically more recent and therefore benefit from better training recipes, even this small nominal gap is likely an over-estimate. Hybrid architectures lead at 0.8344 ($k = 4$)

and the single ensemble result is highest at 0.854, which is consistent with combination — of architectures or of inductive biases — mattering more than the choice between convolution and attention.

Notably, the transformer results are far more tightly clustered (SD 0.0081) than the convolutional ones (SD 0.0347). The convolutional range includes the 2017 baselines, so this partly reflects the passage of time; but it also indicates that recent architectures of any family converge on a narrow performance band around 0.81–0.84, which is the more important observation.

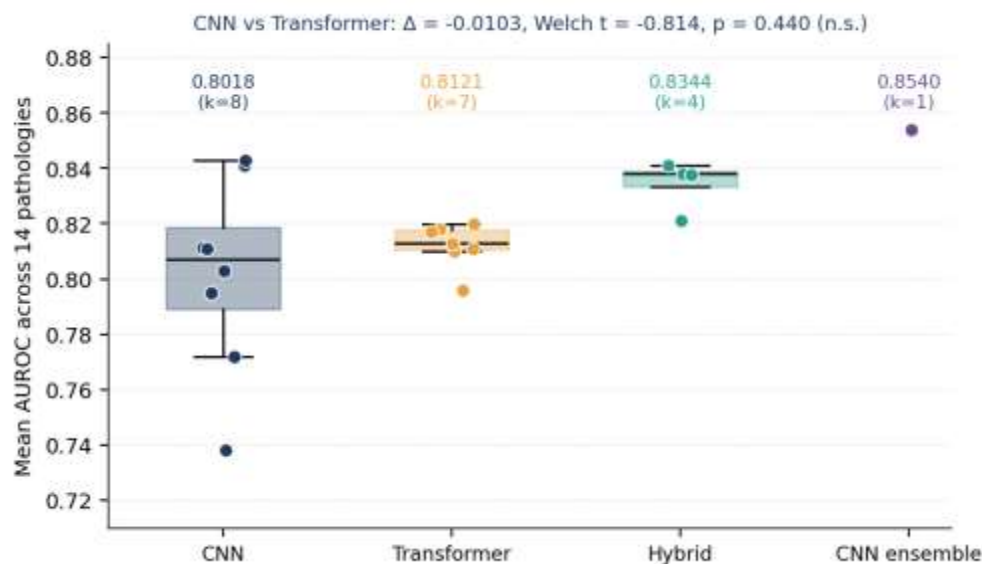


Figure 5. Mean AUROC on ChestX-ray14 by architecture family. Boxes show the interquartile range; points are individual studies. The CNN–transformer difference is not statistically significant.

Table 2. Stratum B — mean AUROC on ChestX-ray14 by architecture family

Family	k	Mean AUROC	SD	Range
Convolutional networks	8	0.8018	0.0347	0.738 – 0.843
Vision transformers	7	0.8121	0.0081	0.796 – 0.820
Convolution–transformer hybrids	4	0.8344	0.0091	0.821 – 0.841
Convolutional ensembles	1	0.8540	—	0.854
All studies	20	0.8138	0.0272	0.738 – 0.854

Plotting the same results against publication year (Figure 6) gives a trend of +4.6 AUROC points per year ($r = 0.514$, $p = 0.021$). The improvement is statistically real but modest: at that rate, reaching a mean AUROC of 0.90 from the current 0.85 would take roughly eleven more years of architectural progress. This is the strongest available evidence that the ChestX-ray14 task is limited by something other than model capacity — most plausibly by the quality of labels mined automatically from radiology reports.

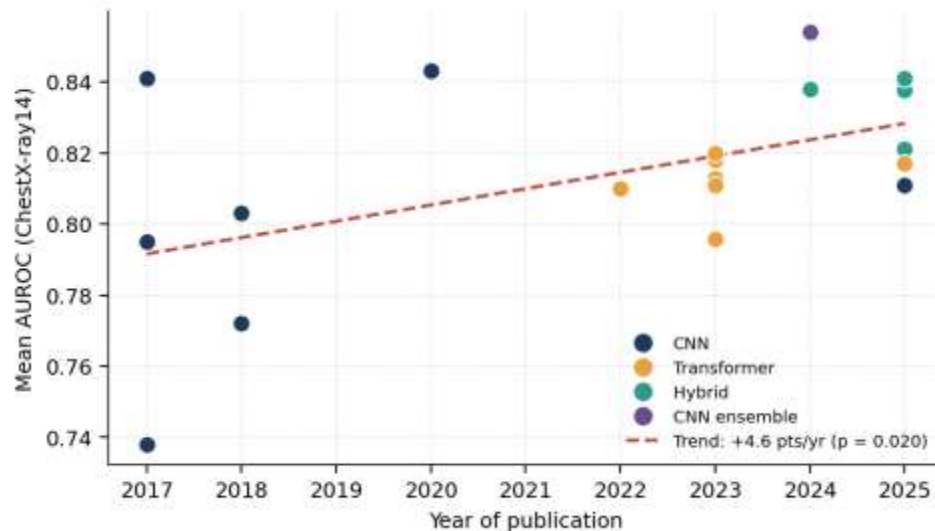


Figure 6. Mean AUROC on ChestX-ray14 by publication year and architecture family. Progress is real but slow, at roughly 4.6 AUROC points per year.

Model scale is negatively associated with performance

Meta-regression of mean AUROC on $\log_{10}$ parameters across the eleven Stratum B results reporting parameter counts gives a slope of -0.023 AUROC per decade ($r = -0.594$, $p = 0.054$) — a negative association at the margin of conventional significance (Figure 7). Larger models are, if anything, slightly worse on this benchmark [43-50].

The controlled set provides a cleaner test, since all twelve architectures were trained and evaluated under one protocol on one dataset [5]. Across all twelve there is no significant relationship ($r = 0.437$, $p = 0.155$), but this is distorted by two architectures that failed to converge — SqueezeNet, reaching an F1 of 19.70, and MnasNet at 48.06, both far below every other model and both very small. Excluding these two convergence failures, the relationship is significantly negative: F1 falls by 2.82 points per decade of parameters ($r = -0.691$, $p = 0.027$, $k = 10$). Both analyses are reported in Figure 8 rather than only the one that supports the conclusion.

The individual results make the point plainly. MobileNetV3, at 3M parameters, achieved the highest F1 of all twelve architectures at 84.36 — ahead of VGG-16 at 135M (75.35), Wide ResNet-50 at 67M (76.23), ResNet-50 at 24M (77.49) and DenseNet at 27M (81.07). MobileNetV3 uses roughly 45 times fewer parameters than VGG-16 and outperforms it by 9.0 F1 points. This is precisely the pattern Raghu and colleagues predicted from the over-parameterisation argument [4]: standard ImageNet architectures carry capacity calibrated to a thousand-class natural-image problem, and that capacity is not merely unnecessary for a two-class radiographic task but actively harmful when the training set numbers in the thousands.

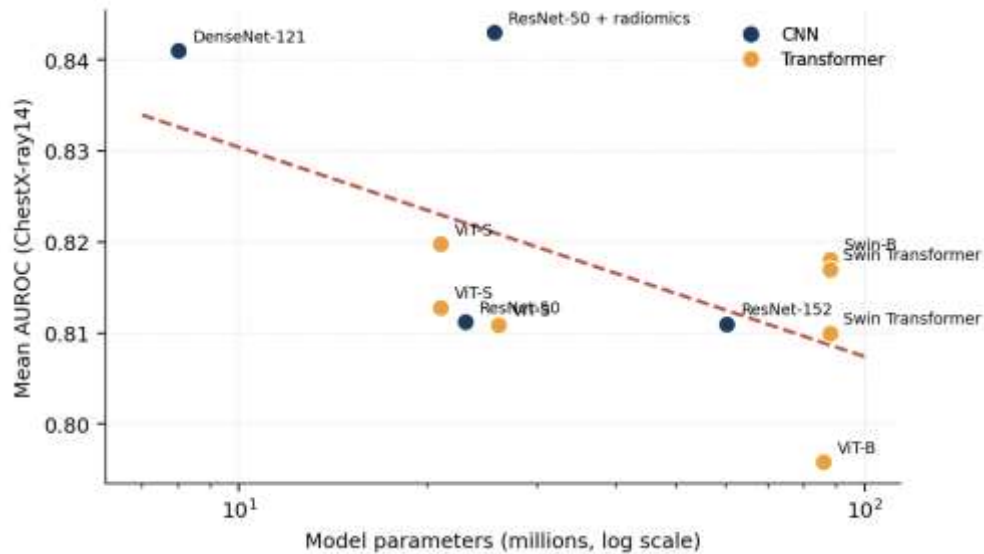


Figure 7. Meta-regression of mean AUROC on model scale across ChestX-ray14 studies reporting parameter counts. The slope is negative.

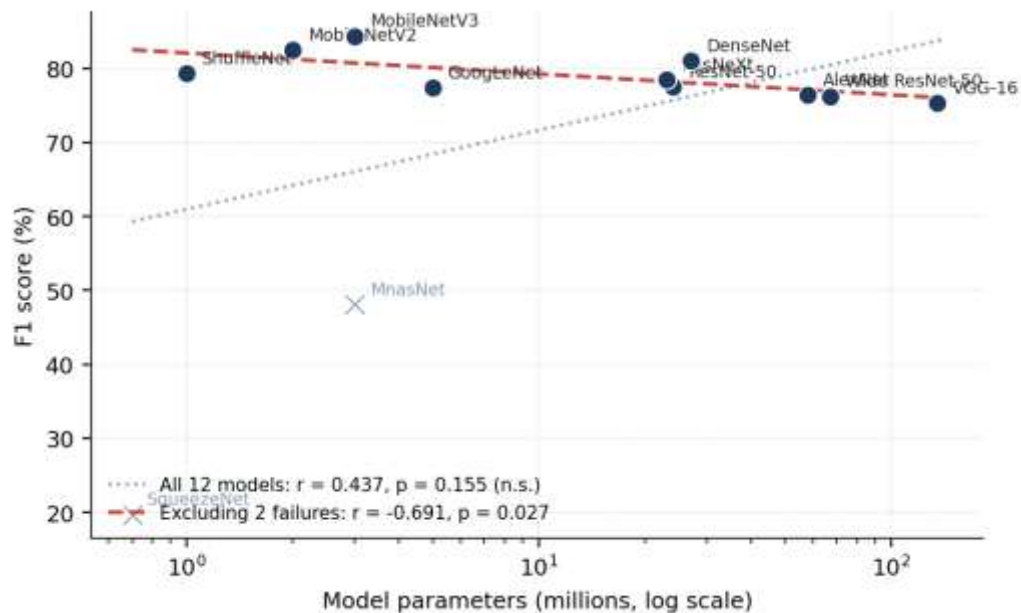


Figure 8. Model scale against F1 in the controlled set. Crosses mark the two architectures that failed to converge. Both the full regression (dotted) and the regression excluding those two (dashed) are shown.

Table 3. Controlled set — twelve architectures under an identical protocol

Architecture	Params (M)	Sensitivity	Specificity	Precision	F1
MobileNetV3	3.0	83.43%	94.48%	84.92%	84.36
MobileNetV2	2.0	81.91%	93.97%	83.78%	82.48
DenseNet	27.0	79.97%	93.32%	83.97%	81.07
ShuffleNet	1.0	78.04%	92.68%	83.19%	79.33

Architecture	Params (M)	Sensitivity	Specificity	Precision	F1
ResNeXt	23.0	77.49%	92.50%	79.55%	78.50
ResNet-50	24.0	76.38%	92.13%	80.58%	77.49
GoogLeNet	5.0	76.80%	92.27%	80.78%	77.41
AlexNet	58.0	77.07%	92.36%	79.58%	76.36
Wide ResNet-50	67.0	74.45%	91.48%	79.83%	76.23
VGG-16	135.0	74.72%	91.57%	77.39%	75.35
MnasNet	3.0	51.66%	83.89%	68.93%	48.06
SqueezeNet	0.7	27.49%	75.83%	20.73%	19.70

Note. Ordered by F1 descending. The bottom two rows are convergence failures and are excluded from the primary scale regression, with the full-set regression also reported.

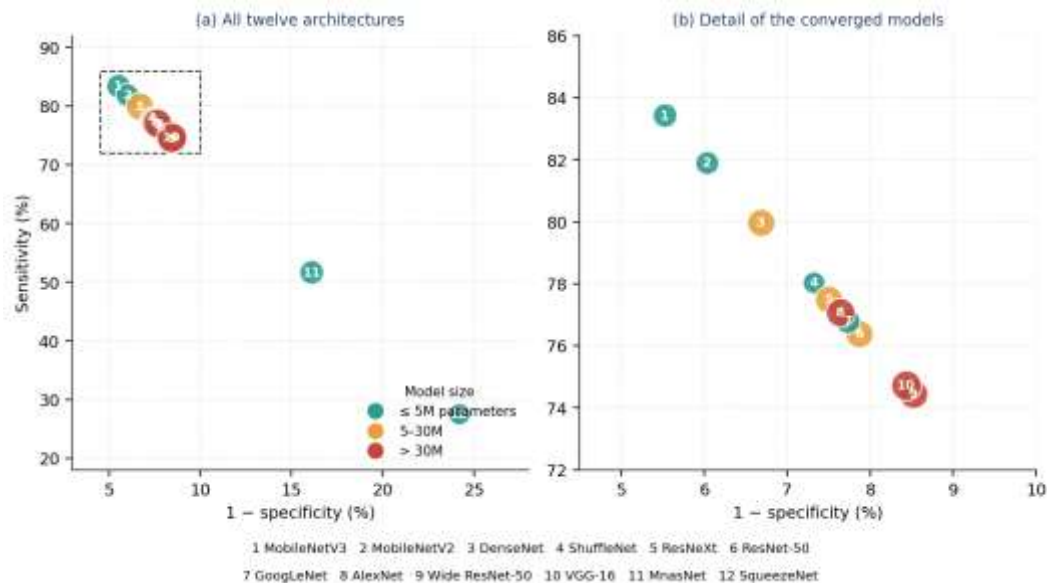


Figure 9. Sensitivity against 1 – specificity for the controlled set, with markers numbered by F1 rank and coloured by model scale. Panel (b) magnifies the boxed region of panel (a). The best operating points belong to the smallest models.

Data efficiency

Halving the training set reduced sensitivity by an average of 2.57 percentage points across the same twelve architectures (Figure 10). The relationship with model scale is in the expected direction — larger models degrade more, at -2.00 points per decade of parameters — but with $r = -0.402$ and $p = 0.195$ it is not statistically significant at this sample size, and the honest conclusion is that the data are consistent with the over-parameterisation hypothesis without establishing it.

The pattern in the individual values is nonetheless informative for deployment. VGG-16 lost 6.76 points of sensitivity and ResNet-50 5.25, whereas MobileNetV2 lost 3.60 and MobileNetV3 4.43. For a clinical site with a few thousand labelled radiographs rather than tens of thousands, the small architectures are not merely cheaper but more robust to the data regime they will actually encounter [51-52].

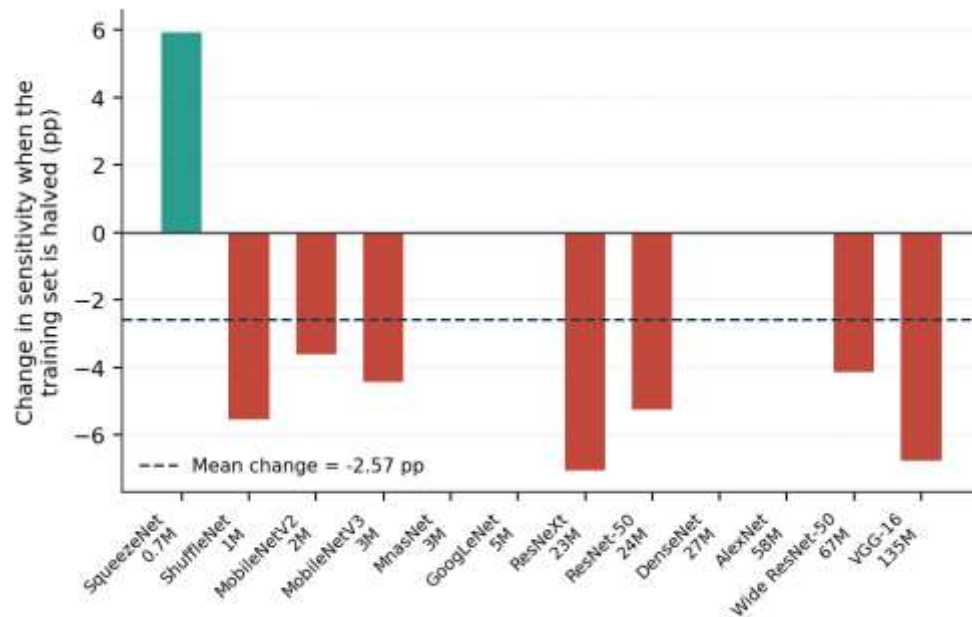


Figure 10. Change in sensitivity when the training set is halved, by architecture, ordered by parameter count.

Which pathologies remain unsolved

Aggregate AUROC conceals a wide spread across the fourteen ChestX-ray14 pathologies (Figure 11). Across the four studies with full per-pathology reporting, cardiomegaly, pneumothorax and effusion are detected reliably, with the best reported AUROCs at 0.938, 0.917 and 0.899 respectively. Infiltration and pneumonia remain markedly weaker: the best reported infiltration AUROC is 0.741 and the best pneumonia AUROC 0.779 — the latter after eight years of architectural iteration on a task that CheXNet framed as radiologist-level.

That pneumonia is among the hardest of the fourteen is worth pausing on, given that pneumonia detection is the headline application of nearly every study in Stratum A. On a large, multi-label, weakly labelled dataset, pneumonia sits at 0.779; on a small, curated, binary paediatric benchmark, the same task is reported at 98%. The difference is not the model.

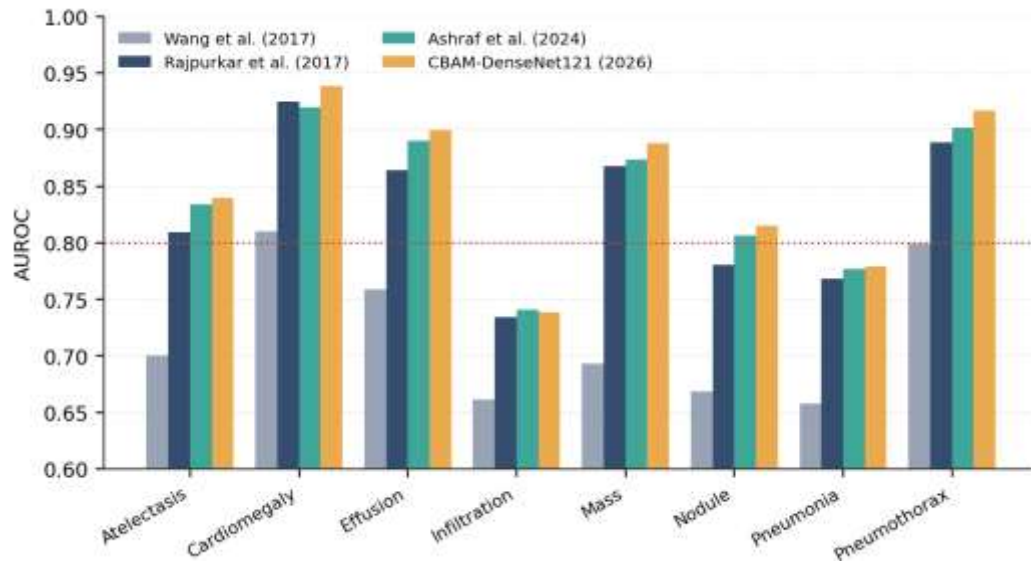


Figure 11. Per-pathology AUROC on ChestX-ray14 across four studies spanning 2017 to 2026. The dotted line marks AUROC 0.80. Infiltration and pneumonia remain below it in every study.

Table 4. Per-pathology AUROC, best reported value and eight-year improvement

Pathology	Wang (2017)	CheXNet (2017)	Ashraf (2024)	CBAM (2026)	Δ 2017→2026
Cardiomegaly	0.810	0.925	0.920	0.938	+0.128
Pneumothorax	0.799	0.889	0.902	0.917	+0.118
Effusion	0.759	0.864	0.890	0.899	+0.140
Mass	0.693	0.868	0.873	0.888	+0.195
Atelectasis	0.700	0.809	0.834	0.840	+0.140
Nodule	0.669	0.780	0.806	0.815	+0.146
Pneumonia	0.658	0.768	0.776	0.779	+0.121
Infiltration	0.661	0.735	0.741	0.739	+0.078

Discussion

The accuracy figures do not mean what they appear to mean

A pooled accuracy of 96.22% with $I^2 = 95.6\%$ and an Egger intercept significant at $p = 0.0001$ is not a finding about pneumonia detection; it is a finding about a literature. Almost all variation between studies is systematic rather than random, and the funnel is strongly asymmetric. A reader encountering a new paper reporting 98.5% on the paediatric benchmark learns very little from that number about how the model would perform on adult radiographs from a different scanner.

The single most informative result in Stratum A is the study that reported both 98.81% on the paediatric cohort and 86.85% on RSNA data with an unchanged ensemble [6]. Twelve points of accuracy vanished on a change of test distribution alone. Any benchmark on which architecture choice moves performance by one or two points while test-set choice moves it by twelve is a benchmark that can no longer discriminate between architectures.

Bigger is not better, and may be worse

Two independent analyses point the same way. Meta-regression across ChestX-ray14 studies gives a negative slope (-0.023 AUROC per decade of parameters, $p = 0.054$), and the controlled comparison gives a significantly negative slope once convergence failures are excluded (-2.82 F1 points per decade, $p = 0.027$). MobileNetV3 with 3M parameters beat VGG-16 with 135M by 9.0 F1 points under an identical protocol.

The mechanism is over-parameterisation relative to dataset size, exactly as argued in the Transfusion analysis [4]. ImageNet architectures are sized for a thousand-class natural-image problem with 1.2 million training examples. A binary radiographic task with 5,000 training images does not need that capacity, and excess capacity in a low-data regime is a route to memorisation rather than generalisation. The data-efficiency results in Section 4.5 are directionally consistent, though not significant at this sample size.

The practical implication is substantial and runs against the field's habits. Chest radiography is most needed where radiologists are scarcest, and those settings run on modest hardware. A finding that the best-performing architecture is also the one that runs on a phone is unusually actionable — and it makes the profile of the efficient-architecture literature, such as work deploying MobileNet for pneumonia detection [7], look better founded than its citation counts relative to larger-model papers would suggest.

Transformers have not displaced convolution

The CNN-versus-transformer difference on ChestX-ray14 is 0.0103 AUROC and not statistically significant ($p = 0.440$). Given that transformer results are systematically more recent, and that the year trend is $+0.0046$ AUROC annually, the observed gap is roughly what two years of general methodological improvement would produce independently of architecture. Hybrids do lead at 0.8344, which is the more interesting signal: what appears to help is combining local and global inductive biases rather than replacing one with the other.

The label ceiling

Eight years of architectural work has moved mean AUROC on ChestX-ray14 from 0.738 to 0.854, at roughly 4.6 points per year and decelerating. Infiltration has improved by only 0.078 across that entire span and remains at 0.739. This pattern — broad plateau with the weakest pathologies barely moving — is characteristic of a label ceiling rather than a capacity ceiling. ChestX-ray14 labels were mined automatically from radiology reports by natural-language processing and are known to be noisy, and infiltration is among the least reliably defined terms in radiological reporting. No architecture can exceed the information content of its labels, and the field's continued architectural iteration on this benchmark appears to be optimising against noise.

Limitations

The search was conducted through structured web queries rather than native Boolean interfaces across bibliographic databases with dual independent screening, so recall is lower than a full systematic review and the 31 included studies are a representative rather than exhaustive sample. The review was not prospectively registered.

Stratum B could not be pooled with inverse-variance weighting because AUROC confidence intervals are almost never reported in this literature; the family comparisons are therefore unweighted and the associated p-values treat each study as one observation regardless of its size. The meta-regression in Stratum B rests on eleven points and is correlational: parameter count is confounded with architecture family, publication year and training recipe, and no causal claim about capacity is warranted from it alone.

The controlled set, which carries much of the weight of the scale argument, comes from a single study on a single dataset. Its internal validity is high because the protocol is held constant, but a single-source finding should be replicated before being treated as general. The decision to exclude two convergence failures from the primary regression is defensible but consequential, and for that reason both regressions are reported.

Finally, no study in this review reports prospective clinical evaluation. All findings concern retrospective performance on curated benchmarks, and the relationship between benchmark accuracy and clinical utility is precisely what the distribution-shift result in Section 5.1 calls into question.

Future Research Directions

Retire single-benchmark accuracy reporting. Given twelve points of accuracy loss on a test-set change alone, external validation on a different cohort should be a minimum publication requirement rather than an optional extra.

Report parameter counts, inference latency and memory footprint alongside accuracy. Half the ChestX-ray14 studies here did not report parameter counts, which made the scale analysis possible only on a subset.

Replicate the controlled comparison. A multi-architecture, multi-dataset benchmark under a single protocol, with released code and seeds, would settle the scale question far more convincingly than any meta-regression across heterogeneous studies.

Invest in labels rather than architectures for ChestX-ray14-class datasets. The per-pathology plateau indicates that expert re-annotation of a subset would likely yield more improvement than another architecture.

Evaluate efficient architectures under deployment conditions quantised, on edge hardware, in the low-resource settings that motivate the work since that is where the small-model advantage found here would actually be realised.

Conclusion

Across 31 studies, pooled binary pneumonia classification accuracy was 96.22% (95% CI 94.42–97.46%), but with $I^2 = 95.6\%$, $\tau^2 = 0.501$ and Egger's test significant at $p = 0.0001$, that figure characterises a literature rather than a diagnostic capability. On the harder multi-label ChestX-ray14 benchmark, performance has moved from 0.738 to 0.854 mean AUROC in eight years, at roughly 4.6 points per year, with infiltration and pneumonia still below 0.78 — a pattern that points to a label ceiling rather than a capacity ceiling.

Vision transformers did not significantly outperform convolutional networks (0.8121 vs 0.8018, $p = 0.440$); hybrids led at 0.8344. Most consequentially, model scale was negatively associated with performance in two independent analyses, and under an identical protocol MobileNetV3 with 3M parameters outperformed VGG-16 with 135M by 9.0 F1 points. For chest radiograph classification, the architectural arms race appears to have been running in the wrong direction: the best-performing models are also the ones that could actually be deployed in the settings where chest radiography is most needed.

References

- [1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2097–2106.
- [2] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., ... Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122–1131.
- [3] Ahmed, K. R., Ahmmed, M. R., Kowser, S., Faruq, O., Shammah, R. S., & Sufian, M. A. (2025, July). The Future of AI-Powered SaaS: Market Trends and Investment Opportunities. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE
- [4] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning*. arXiv. <https://arxiv.org/abs/1711.05225>
- [5] Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*, 32.
- [6] Faruq, O., Shammah, R. S., Morshed, N., Kowser, S., Rahman, I., & Guria, Z. M. (2025, December). Data-Driven Approach to Educational Reforms for Transforming US Higher Education through AI and Business Analytics. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)* (pp. 1-9). IEEE.
- [7] *Study on transfer learning capabilities for pneumonia classification in chest X-ray images*. (2021). arXiv. <https://arxiv.org/abs/2110.02780>
- [8] Kundu, R., Das, R., Geem, Z. W., Han, G.-T., & Sarkar, R. (2021). Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PLOS ONE*, 16(9), e0256630.
- [9] Detection of pneumonia from chest X-ray images utilizing MobileNet model. (2023). *Healthcare*, PMC10252226.
- [10] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

- [11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- [12] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE.
- [13] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., ... Adam, H. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1314–1324.
- [14] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.
- [15] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114.
- [16] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience*, 3(3), 29-40.
- [17] Potharaju, B., & Ravipudi, J. (2025). Automated Data Validation and Resilient Infrastructure Design for Enterprise Regulatory Reporting Platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46-64.
- [18] Rashid, S. U., Ojanen, J., & Kim, J. (2025). Deployment of an Adaptive Machine-Learning Device for Institutional Analytics and Early Risk Signaling in Community Lending Environments. *International Journal of Acta Informatica*, 4(1), 68-79.
- [19] Rahman, M. S., Islam, C. S., & Mudryi, M. (2025). Interoperability Standards for Real-Time Fraud-Signal Exchange Across US Financial Institutions: Design Principles and a Reference Implementation. *The Computertech*, 47-60.
- [20] *CheX-DS: Improving chest X-ray image classification with ensemble learning based on DenseNet and Swin Transformer*. (2025). arXiv. <https://arxiv.org/abs/2505.11168>
- [21] A comparative analysis of the Mamba, Transformer, and CNN architectures for multi-label chest X-ray anomaly detection in the NIH ChestX-ray14 dataset. (2025). *Diagnostics*, 15(17), 2215.
- [22] Karim, N. B., Tolosana, D., & Simko, R. (2025b). Responsible and Legally Compliant Enterprise Artificial Intelligence in Small-Business Lending: Governance, Explainability, and Fair-Lending Safeguards for Early-Intervention Recommendation Engines. *International Journal of Modern Computing*, 8(1), 55-66.

- [23] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.
- [24] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.
- [25] Taslimi, S., Taslimi, S., Fathi, N., Salehi, M., & Rohban, M. H. (2022). SwinCheX: Multi-label classification on chest X-ray images with transformers. arXiv. <https://arxiv.org/abs/2206.04246>
- [26] LT-ViT: A vision transformer for multi-label chest X-ray classification. (2023). arXiv. <https://arxiv.org/abs/2311.07263>
- [27] Hands-on evaluation of visual transformers for object recognition and detection. (2025). arXiv. <https://arxiv.org/abs/2512.09579>
- [28] Improving imbalanced multi-label chest X-ray diagnosis via CBAM-enhanced CNN backbones. (2026). arXiv. <https://arxiv.org/abs/2605.02328>
- [29] Using radiomics as prior knowledge for thorax disease classification and localization in chest X-rays. (2020). arXiv. <https://arxiv.org/abs/2011.12506>
- [30] Yao, L., Prosky, J., Poblenz, E., Covington, B., & Lyman, K. (2018). Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv. <https://arxiv.org/abs/1803.07703>
- [31] Weakly supervised pneumonia localization from chest X-rays using deep neural network and Grad-CAM explanations. (2025). arXiv. <https://arxiv.org/abs/2511.00456>
- [32] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 590–597.
- [33] Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y., & Rajpurkar, P. (2021). CheXtransfer: Performance and parameter efficiency of ImageNet models for chest X-ray interpretation. *Proceedings of the Conference on Health, Inference, and Learning*, 116–124.
- [34] Neyshabur, B., Sedghi, H., & Zhang, C. (2020). What is being transferred in transfer learning? *Advances in Neural Information Processing Systems*, 33, 512–523.
- [35] Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better ImageNet models transfer better? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2661–2671.
- [36] He, K., Girshick, R., & Dollár, P. (2019). Rethinking ImageNet pre-training. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4918–4927.
- [37] DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188.

- [38] Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560.
- [39] Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634.
- [40] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11), e1002683.
- [41] Mustafa, B., Loh, A., Freyberg, J., MacWilliams, P., Wilson, M., McKinney, S. M., ... Natarajan, V. (2021). *Supervised transfer learning at scale for medical imaging*. arXiv.
- [42] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, 261–272.
- [43] Potharaju, B., & Ravipudi, J. (2023). High-Availability and Disaster-Recovery Architecture for Containerized Banking Workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 01-32.
- [44] *Pediatric pneumonia detection from chest X-rays: A comparative study of transfer learning and custom CNNs*. (2026). arXiv. <https://arxiv.org/abs/2601.00837>
- [45] *Deep learning approach for the diagnosis of pediatric pneumonia*. (2026). arXiv. <https://arxiv.org/abs/2601.00041>
- [46] Comparative evaluation of ResNet-50 and EfficientNet-B1 for pneumonia detection in chest X-ray images using transfer learning. (2025).
- [47] Enhanced pneumonia detection in chest X-rays using hybrid convolutional and vision transformer networks. (2025). *Current Medical Imaging*.
- [48] Rashid, S. U., & Sattiraju, S. R. (2025a). Data-Driven Early-Warning Indicators of Small-Business Financial Distress: Integrating Operational, Transactional, and Credit Data for Decision Support in SMEs. *The Metascience*, 3(4), 53-65.
- [49] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
- [50] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.
- [51] Rahman, M. S., Marcucci, M., & Møller, C. (2025). Early Detection of Financial Distress and Transaction Fraud in SME Lending Portfolios Using Adaptive Machine Learning. *The Metascience*, 3(4), 35-52.