

A Comparative Study of Convolutional and Transformer Models for Pneumonia Detection from Chest X-ray Images

with a proposed CNN-Transformer hybrid

PRESENTED BY

Hussam alden Abushalha ID 253002964

Abdulaziz Dhafer Al Obud ID 253001366

SUPERVISOR

Dr. Hagar Salah

Midocean University

Content

01	Problem & gap	Why automated pneumonia screening, and what the literature leaves open	~2 min
02	Method	Dataset, protocol, five baselines and the proposed hybrid	~4 min
03	Results	Overall, per-class, and a significance test of the hybrid	~4 min
04	Deployment	Turning the trained models into a working prototype	~1 min
05	Conclusions	Contributions, limitations, and future work	~2 min

O I

Problem & gap

Why automated pneumonia screening, and what fifteen recent studies leave open.

Chest X-ray reading is the bottleneck, not the imaging

Chest radiography is the primary imaging test for pneumonia because it is fast, widely available and cheap. The constraint is interpretation: the same radiograph can lead two clinicians to different conclusions, and radiologist availability is lowest exactly where the disease burden is highest.

< 5 · > 65

The two age groups with the highest pneumonia mortality

Subjective

Interpretation varies between readers and with fatigue

Scarce

Specialist radiologists are least available in low-resource settings

WHY DEEP LEARNING

Reproducible

the same image always yields the same output

Fast

inference in well under a second per image

Consistent

no fatigue effect across a long reading list

Scalable

extends screening capacity where staff are scarce

Fifteen studies, 2023–2026 — three clear families

All fifteen are summarised in the thesis; the full comparison table is in the backup slides.

CNN & transfer learning	7 STUDIES	Vision transformers	3 STUDIES	Hybrids & ensembles	5 STUDIES
Bundea 2024	97.0%	Singh 2024	97.61%	Boukhe 2025	99.35%
DenseNet variants		ViT (DeiT)		CNN+GRU+SNN+Attn	
Buriboev 2025	98.7%	Shomsomi 2026	98.59%	Mustapha 2025	98.72%
CNN + CLAHE		ViT fine-tuned		CNN + Swin blocks	
Devaux 2024	≈ radiologists	Alghamdi 2024	97.61%	Ridouane 2025	best of 4
EfficientNet-B4		ViT vs CNN ×3 sets		PELM ensemble	
Kareem 2024	93.0%			An 2024	95.19%
Federated CNN				EffNet+DenseNet+Attn	
Mature and cheap to deploy, but plateauing.		Beat CNNs, but degrade across datasets.		Lead the field — but heavy and rarely tested for significance.	

What the literature leaves open — and what we asked

THREE GAPS

I Protocols are not comparable

Reported accuracies come from different splits, preprocessing and augmentation, so architectures cannot be ranked across papers.

2 Gains are rarely tested

Hybrids are declared better on a fraction of a percent, with no significance test on paired predictions.

3 Results stay in notebooks

Very few studies deliver anything a clinician could actually run.

RESEARCH QUESTIONS

- RQ1**

How do leading CNN and transformer architectures compare on pneumonia detection under one identical protocol?
- RQ2**

Does fusing convolutional and transformer features into a hybrid produce a measurably better classifier?

OBJECTIVES

- Fine-tune 5 models**

3 CNNs + 2 transformers, identical conditions
- Design a hybrid**

feature-level fusion, evaluated the same way
- Evaluate rigorously**

accuracy · recall · F1 · AUC · McNemar
- Deliver a prototype**

a local web application serving all six models

02

Method

One dataset, one protocol, five baselines and the proposed hybrid.

Kaggle Chest X-Ray Pneumonia — 5,850 radiographs

CLASS DISTRIBUTION

1,575	4,275	5,850
Normal	Pneumonia	Total
26.9%	73.1%	100%

Paediatric chest radiographs, Guangzhou Women and Children's Medical Centre, labelled Normal / Pneumonia.

SPLIT

70%	10%	20%
Training	Validation	Test
learn parameters	monitor & tune	held out, used once

Split by image, stratified by class. The test set is touched only for the final numbers reported in this talk.

PREPROCESSING

Resize	Normalise	Augment	Leave alone
224 × 224 px	per-channel	rotation · zoom · h-flip	val & test
matches every backbone's input	stabilises and speeds up training	training set only — a regulariser	no augmentation, so evaluation is clean

Class imbalance is real and we did not hide it: pneumonia outnumbered normal roughly 3:1, which is why accuracy alone is not reported — precision, recall, F1 and ROC-AUC are all shown per class.

One pipeline, six models, identical conditions

HELD CONSTANT

Same split	Same 224 × 224 input	Same augmentation
Same optimiser (Adam)	Same loss (BCE)	Same batch size (32)
Same early stopping	Same test set	Same metrics

Any difference in the results is therefore attributable to the architecture, not the setup.

Five pre-trained models, all fine-tuned by transfer learning

Every backbone starts from ImageNet weights; the classifier head is replaced with a two-class head and the network is fine-tuned on the training split.

CONVOLUTIONAL — LOCAL TEXTURE

MobileNetV2

Depthwise-separable convolutions.
Lightweight reference point — what a low-power device could realistically run.

EfficientNetBo

Compound scaling of depth, width and resolution. Strong accuracy per parameter; later becomes the CNN half of the hybrid.

ResNet50

Residual connections. The default backbone in medical imaging, included as the standard of comparison.

TRANSFORMER — GLOBAL CONTEXT

ViT

Splits the image into 16×16 patches and applies self-attention across all of them from the first layer — global context immediately.

Swin

Attention inside shifted local windows, merged hierarchically. Multi-scale like a CNN, at lower cost than full ViT; becomes the transformer half of the hybrid.

The CNN–Transformer hybrid

DESIGN

Parallel, not stacked	Feature-level fusion	Compact head	Two-stage training
both backbones see the same image independently	1,280 (CNN) + 1,024 (Swin) = 2,304-dim vector	Dense 512 + ReLU → dropout → Dense 2 → softmax	head first with backbones frozen, then end-to-end at a low learning rate

Why fuse features, not votes?

Averaging probabilities discards everything the two backbones disagree about. Concatenating the penultimate vectors lets the head learn how to weigh local texture against global context per case.

Why warm-start from fine-tuned checkpoints?

Training 2,304 fused dimensions from scratch on ~4,000 images would overfit. Starting from converged backbones makes the fusion head the only thing that has to be learned.

Why EfficientNetBo + Swin?

The best accuracy-per-parameter CNN paired with the hierarchical transformer — the two strongest complementary encoders from our own baselines.

Configuration held identical across all six models

Environment	Google Colab (GPU)	Optimiser	Adam
Language	Python 3	Loss	Binary cross-entropy
Frameworks	TensorFlow/Keras (CNNs) · PyTorch (transformers)	Batch size	32
Input size	224 × 224 × 3	Epochs	15, with early stopping
Initialisation	ImageNet pre-trained	Test protocol	Held-out set, evaluated once

WHY THIS MATTERS FOR THE ARGUMENT

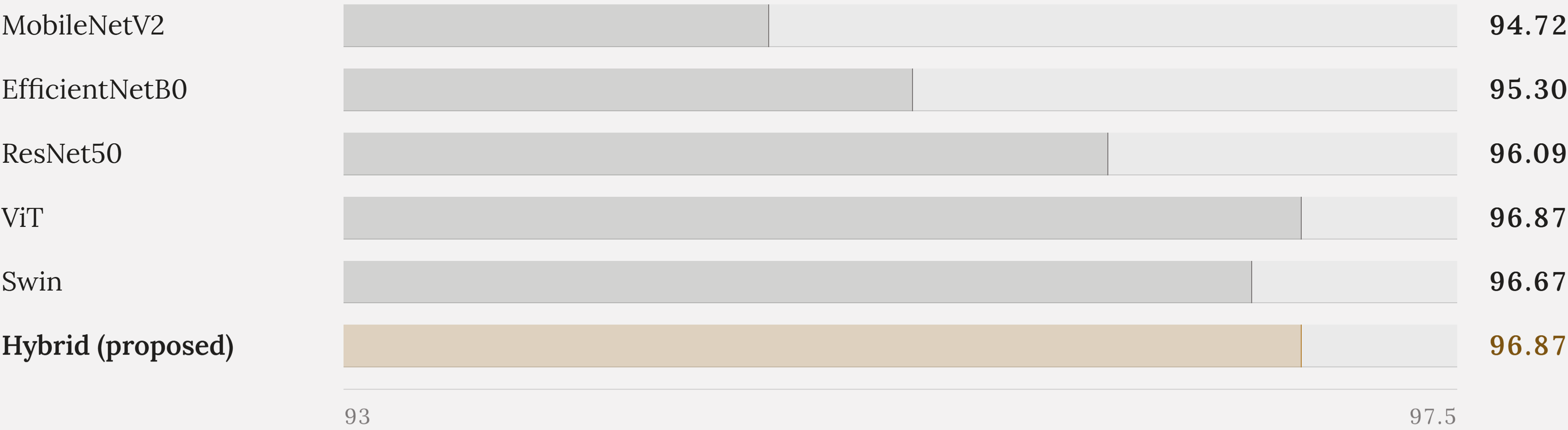
Published pneumonia accuracies range from 93% to 99.35%, but every one comes from a different split, preprocessing chain and augmentation policy — so the spread measures experimental design as much as architecture. Fixing all of it, and varying only the model, is what lets us claim the differences reported next are attributable to the architecture itself.

03

Results

Accuracy, ranking, errors — and a significance test of the hybrid.

Test accuracy — all six models, one protocol



Within 2.2 points

the whole field, from MobileNetV2 to the hybrid

Transformers ahead by 0.78 pt

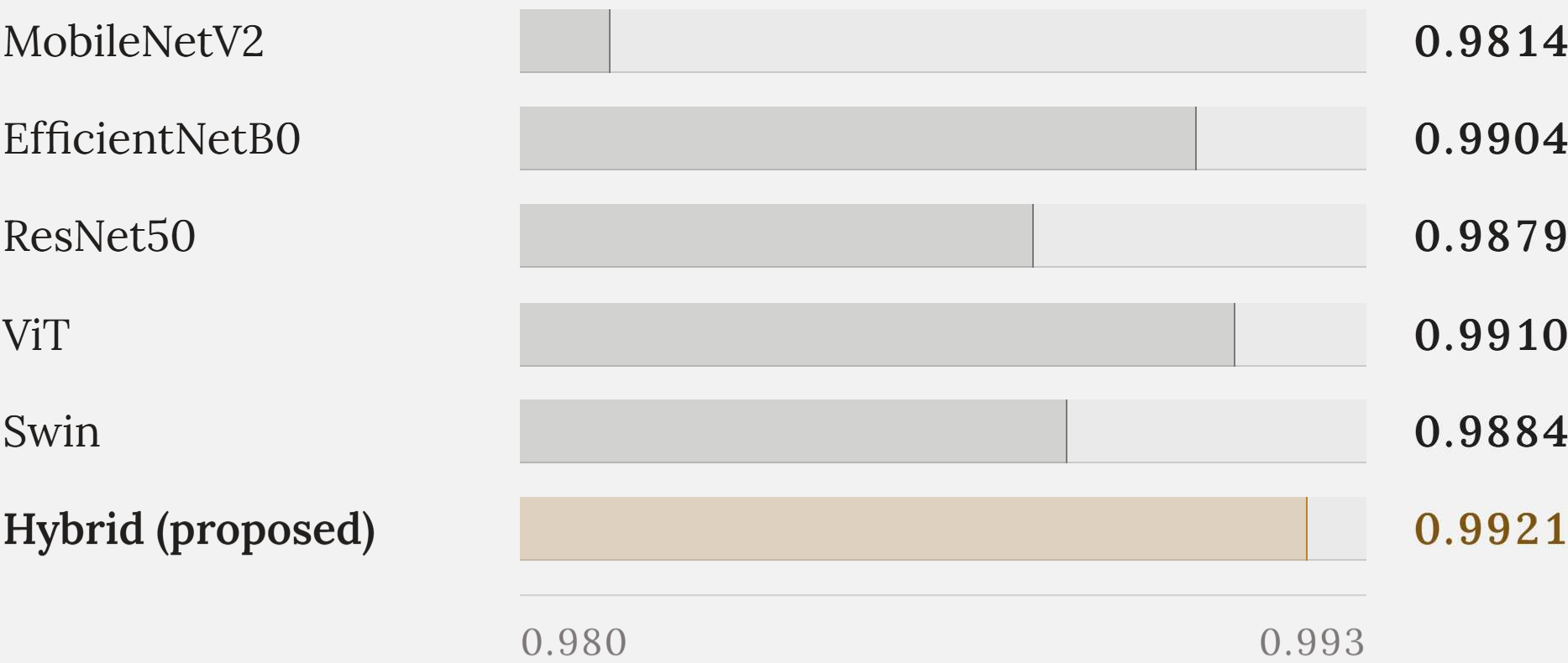
96.87% vs 96.09% for the strongest CNN

Axis truncated at 93%

without truncation the six bars are indistinguishable

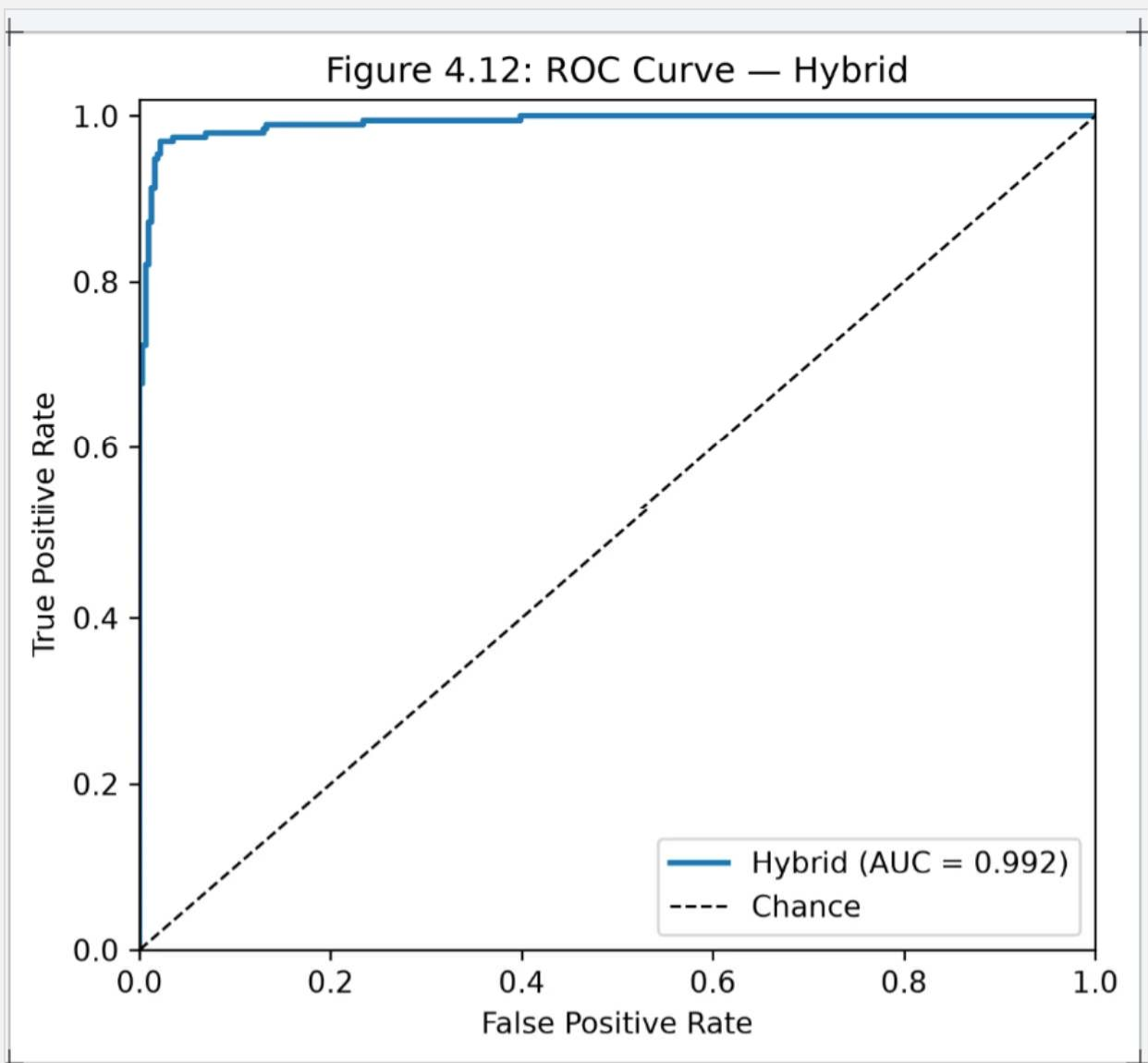
Ranking quality — where the hybrid does win

ROC-AUC BY MODEL



0.9921 — the hybrid takes the highest AUC of the six, on identical predictions to ViT's at the default threshold.

ROC CURVE — PROPOSED HYBRID



AUC integrates every threshold, not just 0.5 — it measures how well a model ranks pneumonia above normal.

Where the hybrid's sixteen errors fall

495

correct of 511

309 normal and 186 pneumonia
films classified correctly

7

false positives

normal films flagged as pneumonia.
In screening this costs a
second read.

9

false negatives

missed pneumonia — clinically the
serious error, and the one to
reduce by lowering the threshold.

0.95

recall on pneumonia

against 0.98 on normal: the harder
class is the one we care about

PER-CLASS SCORES — PROPOSED HYBRID

Normal precision 0.97 · recall 0.98 · F1 0.97

Pneumonia precision 0.96 · recall 0.95 · F1 0.96

All six models' per-class scores are in backup B3.

THE THRESHOLD IS A CLINICAL CHOICE, NOT A DEFAULT

At 0.5 the model balances the two error types. A screening deployment would move the operating point to favour sensitivity, accepting more second reads to miss fewer cases — which is exactly what the hybrid's higher AUC buys.

Is the hybrid actually better? We tested it.

Almost no study in our review checks whether its reported improvement is statistically real. We applied McNemar's test to the paired predictions of the hybrid and the best single model on the same held-out images.

507

images classified identically

the two models agree on 99.2% of the test set

4

images where they disagree

the entire basis for any claim of superiority

$\chi^2 = 0.25$

McNemar test statistic

1 degree of freedom, continuity-corrected

$p = 0.617$

not significant at $\alpha = 0.05$

we cannot reject the null of equal performance

The hybrid matches the best single model — it does not significantly beat it.

WHAT WE CAN CLAIM

Feature fusion produces a classifier at least as good as the strongest individual model, with the highest AUC of the six — better ranking of cases across thresholds.

WHAT WE CANNOT CLAIM

A significant accuracy gain. Reporting one on a 0.2-point difference, without a paired test, would not be defensible.

WHY THIS IS STILL A RESULT

It maps the ceiling. On a single-source benchmark where baselines already exceed 96%, there is little headroom left for fusion to exploit.

O4

Deployment

Turning six trained models into something a clinician could run.

From notebook to a working prototype

WHAT IT DOES

Serves all six trained models from one local interface

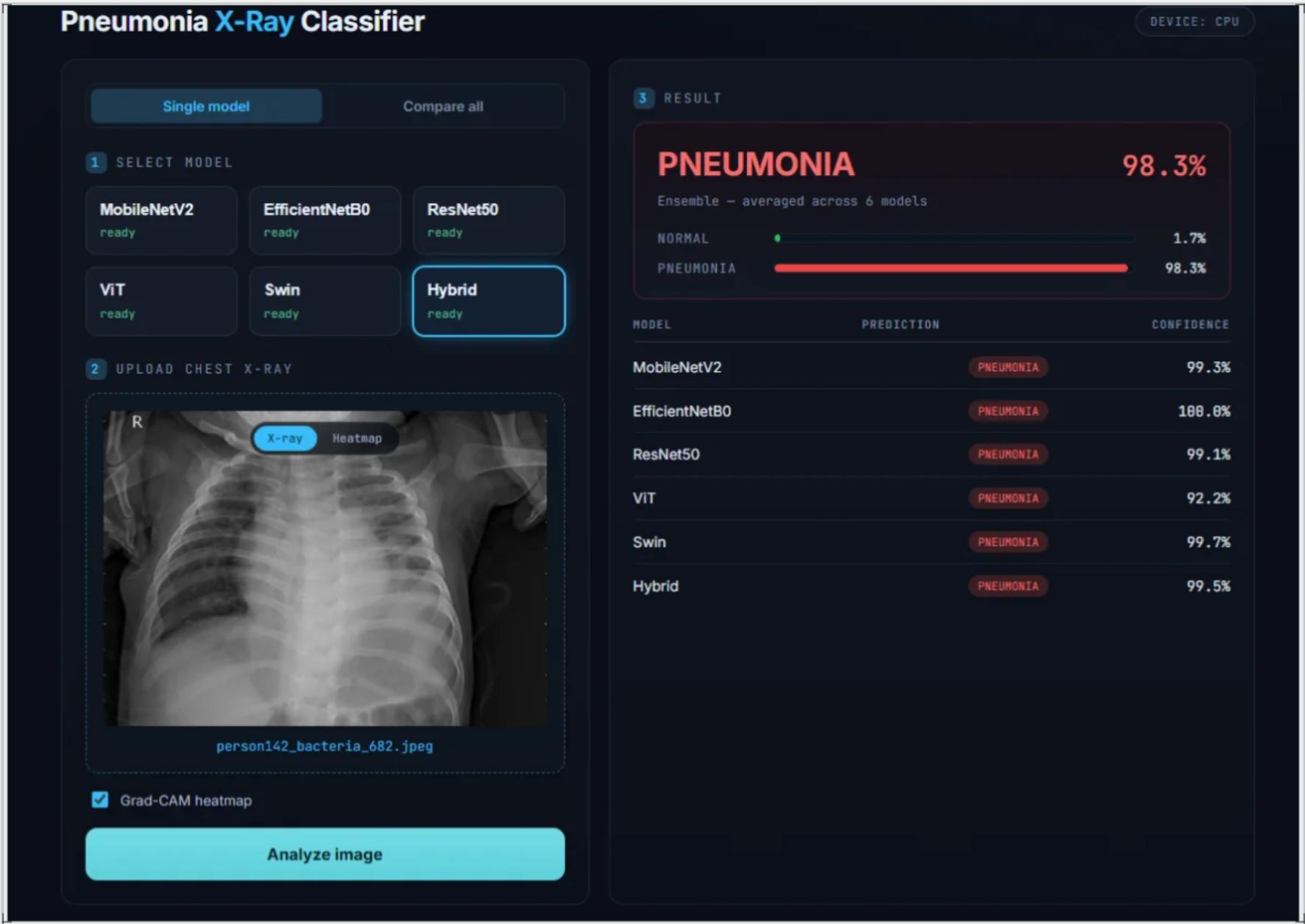
Upload an X-ray, get a class, a confidence, and a per-class probability breakdown

Reports inference time — under a second on CPU

Weights load on demand and stay cached, so switching models is instant

NOT A MEDICAL DEVICE

A research and educational prototype. Clinical use would require prospective validation and regulatory approval, and every prediction must be confirmed by a qualified radiologist.



05

Discussion & conclusions

How this sits against the literature, what it cannot show, and what comes next.

How this sits against the literature — and what it cannot show

IN LINE WITH THE FIELD

Transformers > CNNs

Our 96.87% vs 96.09% reproduces the ordering reported by Singh (2024) and Shomsomi (2026).

Below the headline numbers

Studies reporting 98–99% use different splits and heavier preprocessing; our single fixed protocol is stricter.

Fusion helps ranking, not accuracy

Consistent with Ridouane's PELM, where the ensemble's advantage was also strongest in ranking quality.

LIMITATIONS WE ACKNOWLEDGE

Single source

One paediatric dataset from one hospital. Cross-dataset generalisation is untested — and Alghamdi (2024) showed that is exactly where accuracy collapses.

No explainability yet

The models give a class and a confidence, but not the region that drove the decision — a barrier to clinical trust.

Binary only

Normal vs pneumonia. No bacterial/viral distinction, no other lung pathology, no comorbidity.

Small test set

At $n \approx 500$, differences under one percentage point are not resolvable, as the McNemar result showed.

Contributions and where this goes next

FOUR CONTRIBUTIONS

I A controlled comparison

Five architectures under one identical protocol, so the ranking reflects the models rather than the experimental setup.

2 A CNN–transformer hybrid

Feature-level fusion of EfficientNetB0 and Swin, matching the best single model and taking the highest AUC at 0.9921.

3 An honest significance test

McNemar on paired predictions, reported as a null result — rare in this literature.

4 A working prototype

A local application serving all six models with sub-second CPU inference.

FUTURE WORK

Multi-centre validation

test on CheXpert, MIMIC-CXR and RSNA to measure the cross-dataset drop directly

Explainability

Grad-CAM or SHAP overlays so a clinician can see the evidence behind a prediction

Threshold calibration

tune the operating point for sensitivity, which is where the hybrid's AUC advantage should pay off

Broader task

bacterial vs viral, and multi-disease screening

Thank you

Questions and discussion

Hussam alden Abushalha · Abdulaziz Dhafer Al Obud
Supervisor: Dr. Hagar Salah · College of Informatics, Midocean
University

Backup slides follow: full literature table · all metrics · per-class scores ·
confusion matrices · training curves · key references

Full comparative summary of reviewed studies

Reported accuracies are not directly comparable: each study uses its own split, preprocessing and augmentation. This is the motivation for our fixed protocol.

STUDY	BEST MODEL	DATASET(S)	ACCURACY
Singh (2024)	ViT (DeiT-Base)	Kaggle CXR	97.61%
Mustapha (2025)	Hybrid CNN + Swin	Kaggle CXR	98.72%
Shomsomi (2026)	ViT, fine-tuned	Kaggle CXR	98.59%
Kareem (2024)	Federated CNN	Multi-institutional (private)	93.0%
Alghamdi (2024)	ViT	Kaggle · BRAX · CheXpert	97.61%
Boukhe (2025)	CNN + GRU + SNN + attention	Kaggle CXR	99.35%
Ridouane (2025)	PELM feature-level ensemble	NIH · CheXpert · PadChest · Kaggle	best of 4
Devaux (2024)	EfficientNet-B4	CheXpert · MIMIC-CXR · ChestX-ray14	≈ radiologists
Buriboev (2025)	CNN + modified CLAHE	Kaggle CXR	98.7%
Saber (2025)	TransUNet + ResNet	Kermany · Cohen	96.04%
Bundea (2024)	DenseNet121/169/201	Kaggle CXR	97.0%
An (2024)	EfficientNetB0 + DenseNet121 + attention	Kaggle CXR	95.19%
Usman (2025)	ResNet-50 (transfer learning)	RSNA	98.0%
Ahmed (2023)	Deep CNN, multi-disease	Kaggle CXR	98.7%
Al-Reshan (2023)	MobileNet	Kaggle CXR	≈95%

All metrics, all six models

APPROACH	MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE	ROC-AUC
Pre-trained CNN	MobileNetV2	0.9472	0.9471	0.9472	0.9470	0.9814
	EfficientNetB0	0.9530	0.9530	0.9530	0.9530	0.9904
	ResNet50	0.9609	0.9610	0.9609	0.9607	0.9879
Transformer	ViT	0.9687	0.9687	0.9687	0.9687	0.9910
	Swin	0.9667	0.9667	0.9667	0.9667	0.9884
Proposed	Hybrid (EffNetB0 + Swin)	0.9687	0.9687	0.9687	0.9687	0.9921

Per-class precision, recall and F₁ — all six models

MODEL	CLASS	PRECISION	RECALL	F1-SCORE
MobileNetV2	Normal	0.95	0.97	0.96
	Pneumonia	0.94	0.92	0.93
EfficientNetB0	Normal	0.96	0.97	0.96
	Pneumonia	0.94	0.93	0.94
ResNet50	Normal	0.96	0.98	0.97
	Pneumonia	0.97	0.93	0.95
ViT	Normal	0.97	0.98	0.97
	Pneumonia	0.96	0.95	0.96
Swin	Normal	0.97	0.97	0.97
	Pneumonia	0.96	0.95	0.96
Hybrid	Normal	0.97	0.98	0.97
	Pneumonia	0.96	0.95	0.96

READING THE HYBRID'S ERRORS

7
false positives
normal films flagged as pneumonia. In screening this costs a second read.

9
false negatives
missed pneumonia. Clinically the more serious error, and the one to reduce by lowering the decision threshold.

Confusion matrices — all six models

ERROR COUNTS

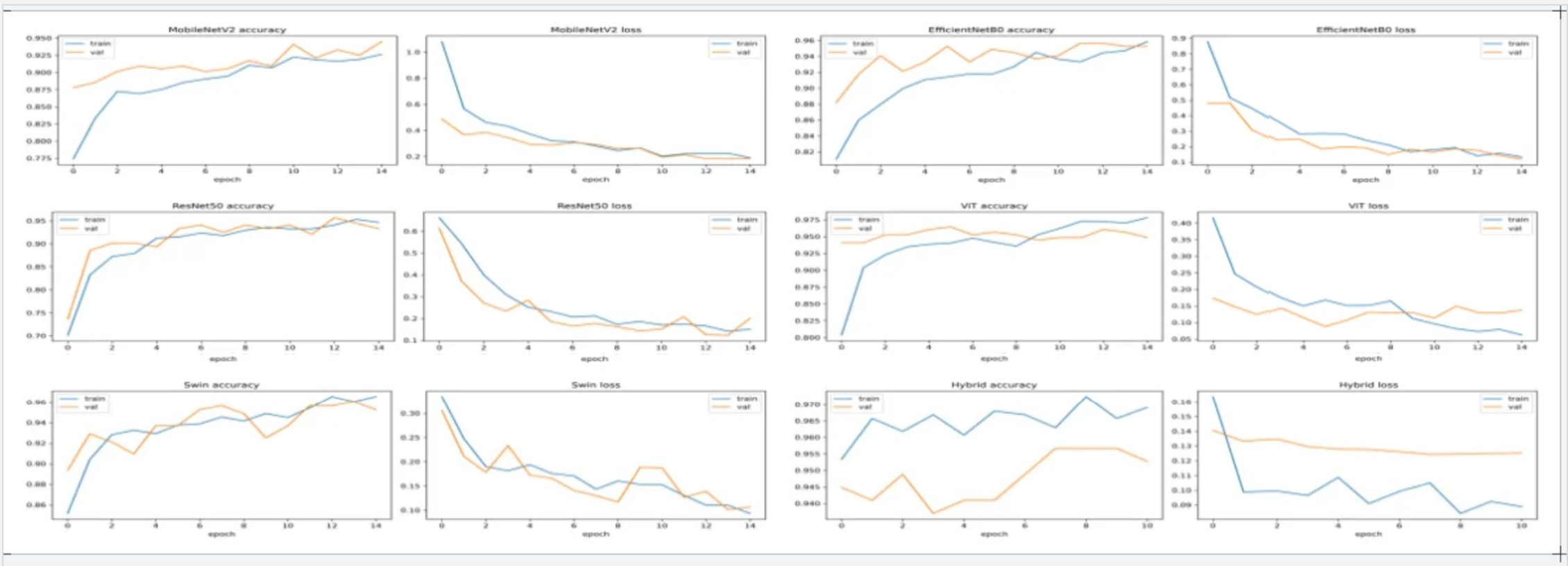
MobileNetV2	11 FP · 16 FN	27
EfficientNetB0	11 FP · 13 FN	24
ResNet50	6 FP · 14 FN	20
ViT	7 FP · 9 FN	16
Swin	8 FP · 9 FN	17
Hybrid	7 FP · 9 FN	16

FP = normal called pneumonia · FN = pneumonia missed · right column = total errors

NOTE

ViT and the hybrid produce identical confusion matrices (309/7/9/186). The hybrid was therefore compared against Swin in the McNemar test, since a test against an identical prediction vector is degenerate.

Training and validation curves — no overfitting



Training and validation curves track each other and converge without a widening gap in every case. All models start high in epoch 1 because the backbones are ImageNet-initialised. The hybrid trains in two stages — fusion head first, then end-to-end — and converges within 11 epochs.

Key references

[1] Singh, S. et al. (2024). Efficient pneumonia detection using Vision Transformers on chest X-rays. Scientific Reports, 14, 2487.

[2] Mustapha, B. et al. (2025). Enhanced pneumonia detection in chest X-rays using hybrid convolutional and vision transformer networks. Current Medical Imaging.

[3] Shomsomi, M., Triawan, W. & Purwadi (2026). Classification of pneumonia using CNN and ViT. JAIEA, 5(2), 2495–2504.

[5] Alghamdi, H. S. et al. (2024). Pneumonia detection from CXR images using deep learning and transfer learning. Biomedical Engineering Letters.

[6] Boukhe, M. et al. (2025). Trustworthy pneumonia detection through attention-guided deep learning. Scientific Reports.

[7] Ridouane, F. L. et al. (2025). PELM: a deep learning model for early detection of pneumonia. Applied Sciences, 15(12), 6487.

[8] Devaux, M. et al. (2024). Validating the accuracy of deep learning for the diagnosis of pneumonia on chest X-ray. European Radiology Experimental, 8, 20.

[9] Buriboev, A. S. et al. (2025). Binary classification of pneumonia in chest X-ray images using CNNs. Sensors, 25(13), 3976.

[16] LeCun, Y. et al. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE.

[18] He, K. et al. (2016). Deep Residual Learning for Image Recognition. CVPR.

[19] Tan, M. & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for CNNs. ICML.

[20] Dosovitskiy, A. et al. (2020). An Image is Worth 16×16 Words. arXiv:2010.11929.

[27] Liu, Z. et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV.

[25] World Health Organization (2023). Pneumonia — fact sheet.

Full reference list (27 entries) in the thesis document.