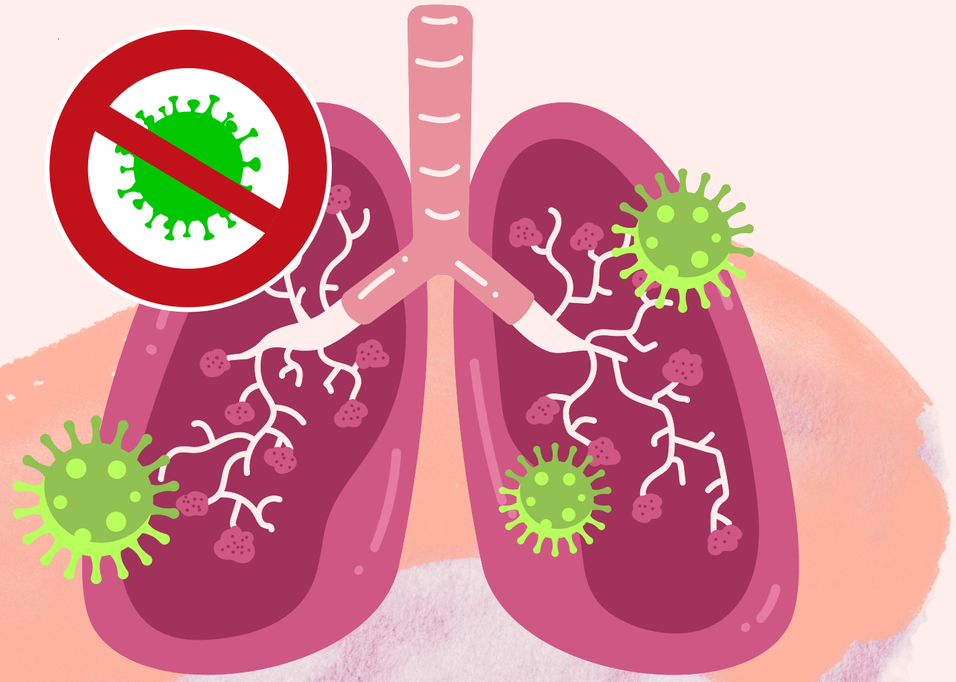
universidade de aveiro
theoria poiesis praxis

deti universidade de aveiro
departamento de eletrónica, telecomunicações e informática

PNEUMONIA Classification



from Pediatric Chest X-Rays with Lightweight Deep Learning Models

Presented by :

Carolina Reis & Jakub Błaszczyk

University of Aveiro, DETI Aveiro, Portugal

Lodz University of Technology, Łódź, Poland

Table of Contents

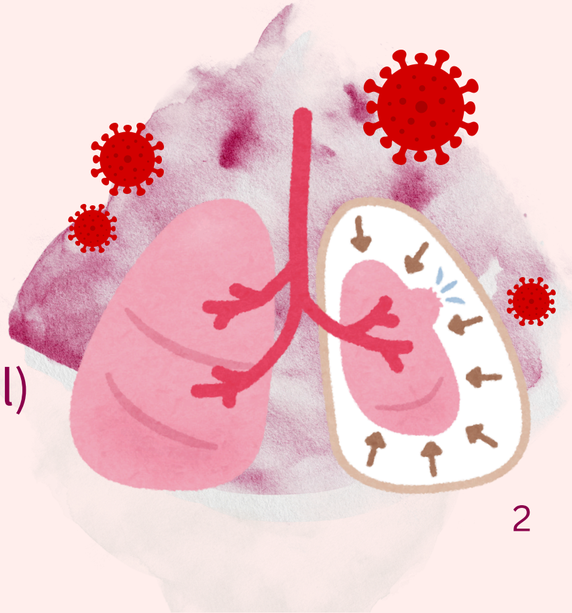
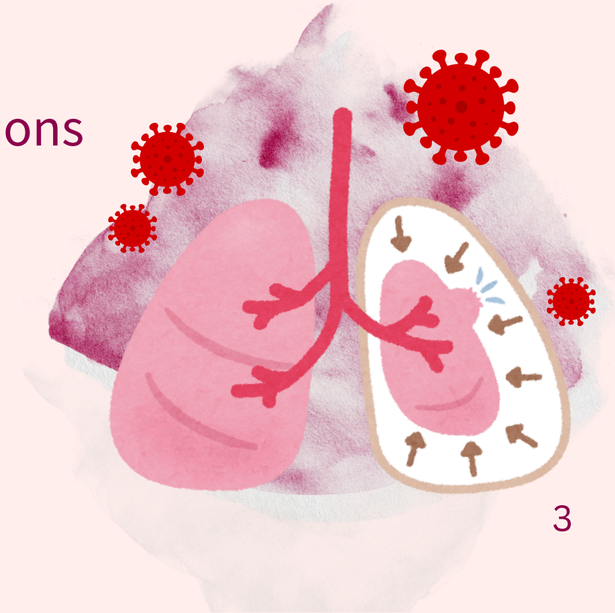
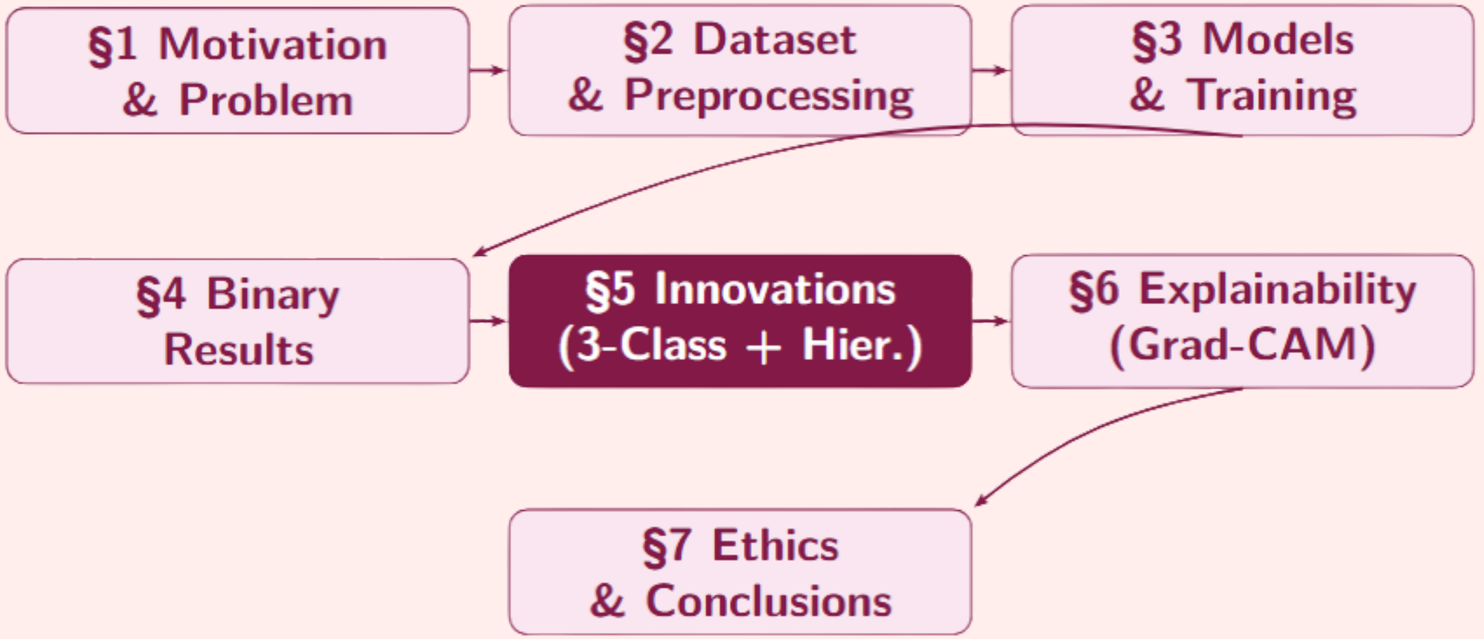


Table of Contents



Outline


§1 Motivation & Problem
§2 Dataset & Preprocessing
§3 Models & Training
§4 Binary Results
§5 Innovations (3-Class + Hier.)
§6 Explainability (Grad-CAM)
§7 Ethics & Conclusions

Motivation & Problem Statement

Fig. 1 - Left: Normal, Center: Bacterial Pneumonia, Right: Viral Pneumonia

Scope

In scope: binary + 3-class classification, Grad-CAM inspection, ensemble, calibration

Out of scope: segmentation, clinical deployment

2.5M+ deaths/year

#1 cause children <5

X-ray most accessible

The Clinical Gap

▶ Diagnosis requires trained radiologists → scarce in low-resource settings

▶ Late or missed diagnosis ⇒ preventable deaths

Our Approach

▶ Baseline: binary Normal vs. Pneumonia

▶ Innovation: 3-class Normal / Bacteria / Virus

▶ 3 models: Custom CNN, ResNet18, DenseNet121

▶ Ensemble, calibration, Grad-CAM

Dataset: Kermany Chest X-Ray Images

Source & Structure

- ▶ Kermany et al. (2018) → public on Kaggle
- ▶ Guangzhou Women & Children's Medical Center
- ▶ **5,863 pediatric X-rays:** NORMAL / PNEUMONIA (binary); filenames encode BACTERIA / VIRUS ⇒ 3-class labels

Split	Normal	Pneumonia	Total
Train	1,341	3,875	5,216
Validation	8	8	16
Test	234	390	624
Total	1,583	4,273	5,856

Table 1 - Binary class distribution across train/val/test splits (Kermany dataset)

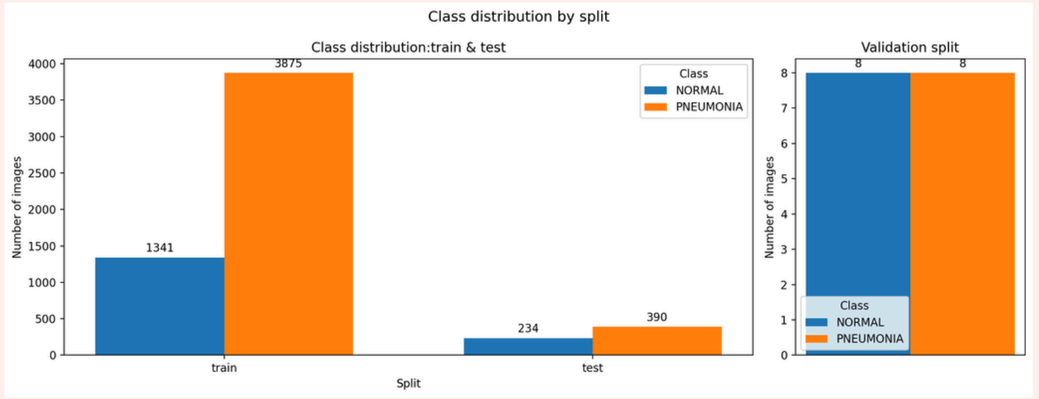


Fig. 2 - Class distribution across splits

Class Imbalance & 3-Class Labels:

Train = **74% pneumonia** vs. 26% normal.
Filenames encode **bacteria** (2,530) vs. **virus** (1,345) ⇒ free 3-class labels.
Recall and AUROC matter more than accuracy.

Preprocessing & Data Augmentation

Applied to All Splits

1. Resize to 224 × 224 px (ResNet/DenseNet standard)

2. Grayscale → RGB: channel triplication

3. Normalize with ImageNet statistics:
μ = [0.485, 0.456, 0.406]
σ = [0.229, 0.224, 0.225]

Train Split Only → Augmentation

▶ Random horizontal flip
▶ Random rotation ±10°

Light augmentation: does not distort pathological patterns (opacities, consolidations).

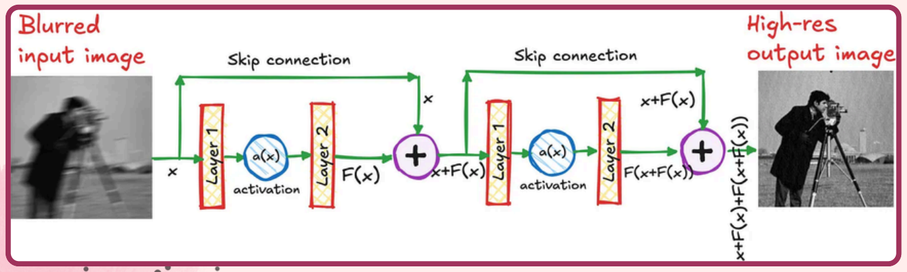


Fig. 3 - ResNet18

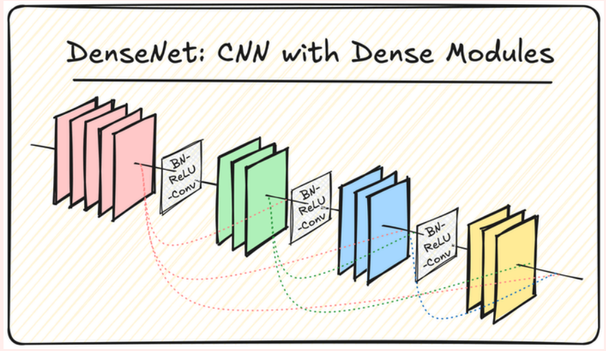
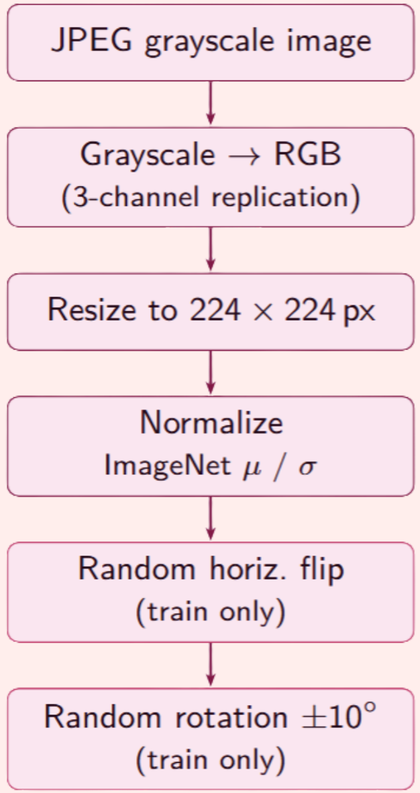


Fig. 4 - DenseNet121



Pipeline & Model Architectures

Why DenseNet121?

CheXNet (Rajpurkar et al., 2017) used DenseNet121 for radiologist-level chest X-ray performance. Dense connectivity encourages feature reuse and excels at detecting subtle diffuse opacities. Fewer params (7M) than ResNet18 (11M) with richer representations.

Custom CNN (Baseline)

3×(Conv+BN+ReLU+Pool)
↓ Global Avg Pool
↓ FC (1) + Sigmoid
No pretrained weights
≈0.5M parameters

ResNet18

18-layer residual net
Residual skip connections
↓ FC head replaced
ImageNet pretrained
Transfer learning
≈11M parameters

Custom CNN (Baseline)

121-layer dense network
Dense feature reuse
↓ Classifier replaced
ImageNet pretrained
CheXNet heritage
≈7M parameters

Training Setup

Shared Hyperparameters

▶ Optimizer: Adam, lr = 10−4

▶ LR scheduler: ReduceOnPlateau (factor 0.5, patience 3)

▶ Loss: BCE (binary) / cross-entropy (3-class)

▶ Batch size: 32 · Early stopping: patience 5

▶ Max epochs: 30 · 3 random seeds

Evaluation Protocol

▶ Fixed train/val/test split → no data leakage

▶ Hyperparams selected on val set only

▶ Test set evaluated once at the end

▶ Report mean ± std over 3 seeds

	CNN	ResNet18	Densenet121
Pretrained	×	✓	✓
Skip / Dense	×	skip	dense
CheXNet	×	×	✓
Params	0.5M	11M	7M

Table 2 - Architecture Comparison

Two-Phase Transfer Learning Strategy

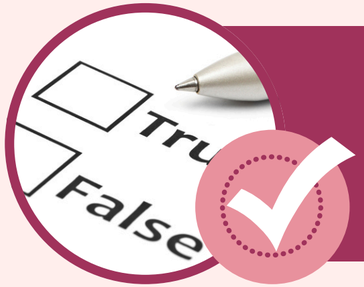
Phase 1 (5 epochs) → Head only
Backbone frozen; only FC head trained

Phase 2 (up to 30 epochs) → Full fine-tune
Entire network unfrozen; LR scheduler active

Quantitative Results: Mean ± Std (3 seeds)

Model	Accuracy	Precision	Recall*	F1	AUROC	ms/img
Custom CNN	0.756±.090	0.768±.110	0.914±.076	0.828±.040	0.886±.016	14.9±3.3
ResNet18	0.830±.018	0.791±.021	0.991±.005	0.879±.011	0.949±.005	13.4±0.8
DenseNet121	**0.841±.020**	**0.799±.020**	**0.997±.002**	**0.887±.012**	**0.966±.007**	**13.8±0.9**

Table 3 - Binary classification test-set results: mean ± std across 3 random seeds
*Recall = pneumonia-class recall (sensitivity).



Clinical Priority: Recall
A false negative (missed pneumonia) is far more costly than a false alarm.
Success criterion: recall > 0.90 ✓
DenseNet121: only 1–2 cases missed per run (out of 390)

Key Observations

- ▶ Transfer learning ≫ baseline on all metrics
- ▶ DenseNet121: recall = 0.997 → near-perfect sensitivity
- ▶ DenseNet121: AUROC = 0.966 (CheXNet-14 reference: 0.927)
- ▶ Inference cost identical across all three models (~13–15 ms)

Visual Comparison of Models

Recall–Specificity Trade-off

Maximising recall necessarily reduces specificity. Some healthy cases are flagged as pneumonia → acceptable in triage: a clinician reviews flagged results. CNN specificity (0.49 ± 0.36) is unstable across seeds → reflects training-from-scratch variance.

Reading the Chart

- ▶ F1/Accuracy: Dense ≈ ResNet ≫ CNN
- ▶ Recall: all > 0.90; TL models near 1.0
- ▶ AUROC: large gap → TL vs. baseline
- ▶ Specificity: moderate (0.49–0.58) for all

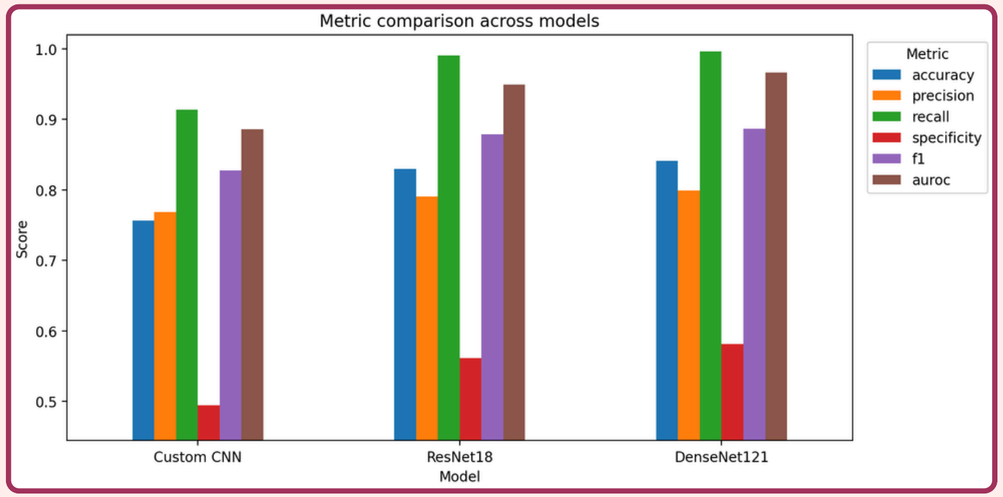


Fig. 5 - Metric comparison across models (binary)

Training Curves: DenseNet121 (Best Model)

Phase 1 → Head Only (5 epochs)

► Rapid loss drop; ImageNet features give a strong starting point

► Fast convergence before unfreezing

Phase 2 → Full Fine-Tune

► Gradual, stable improvement

► LR scheduler reduces on plateau

► Early stopping prevents overfitting

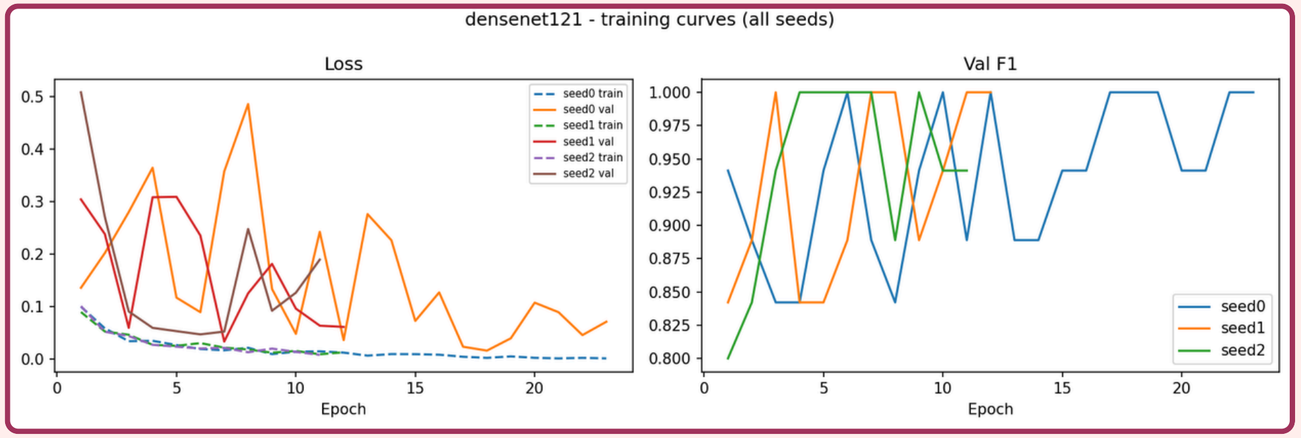


Fig. 6 - Densenet121 training curves

Stability Across Seeds

Std ≈ 0.020 on all metrics across 3 seeds → low variance, robust procedure.

Confusion Matrix: Binary Ensemble (6 checkpoints)

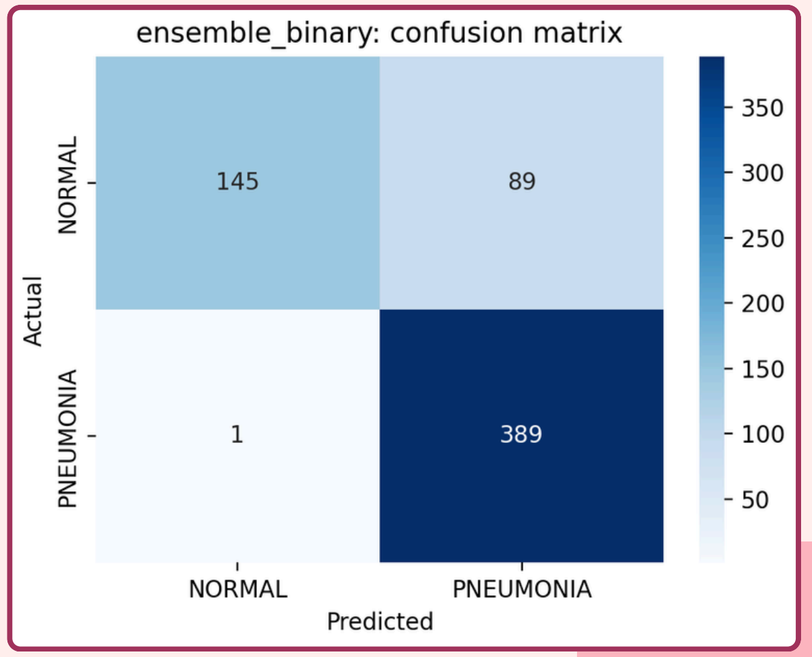


Fig. 7 - Ensemble binary confusion matrix

	Pred. Normal	Pred. Pneumonia
True Normal	TN = 145	FP = 89
True Pneumonia	FN = 1	TP = 389

Table 4 - Test Set Breakdown (624 images)

Recall = 389/390 = 0.997
Specificity = 145/234 = 0.620
Precision = 389/478 = 0.814

Clinical Interpretation

Only 1 pneumonia cases missed out of 390.
89 false alarms ⇒ flagged for clinician review, not autonomous decision.
This reflects the correct triage role of the system.

Innovation I: Three-Class Classification

Why split pneumonia?

- ▶ Bacterial vs. viral pneumonia need different treatment (antibiotics vs. supportive care)
- ▶ Aetiology is encoded in the filenames (_bacteria_ / _virus_)
- ▶ Labels derived automatically → no manual annotation

Model	Accuracy	$F1_M$	AUROC
ResNet18	0.769	0.751	0.944
DenseNet121	0.799	0.784	0.944
Dense121 (hier.)	0.790	0.774	0.944
Ensemble	**0.822**	**0.808**	**0.960**

Table 5 - Patient-aware split (leakage-free). Hierarchical = two-stage; does not beat the flat softmax.

	Normal	Bacteria	Virus
F1	0.78	**0.92**	0.72
Recall	0.65	0.97	0.85

Table 6 - Per-Class F1 (Ensemble)

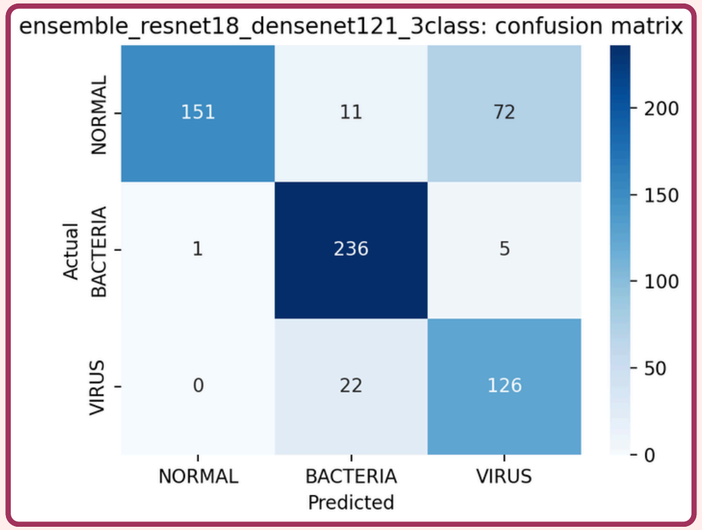


Fig. 8 - Ensemble 3-class confusion matrix

Virus is the hardest class; the main error is Normal → Virus (faint infiltrates).

Innovation II: Hierarchical Two-Stage Classifier

Motivation

The original paper never trained a flat 3-way classifier — it used two separate binary classifiers. We replicate that philosophy and compare head-to-head.

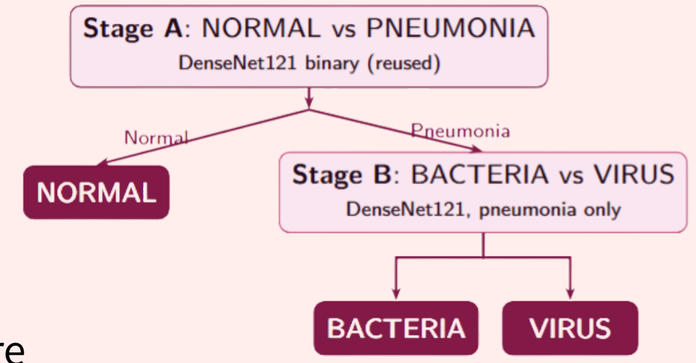
Why it doesn't win?

▶ Stage A errors propagate to Stage B (no recovery)

▶ Normal → Virus confusion unchanged by the cascade

▶ VIRUS bottleneck is intrinsic to the data

▶ Flat softmax weighs all 3 classes jointly, more robust ⇒ Negative but informative: VIRUS is a genuine data-level bottleneck.



	Flat	Hier.
Accuracy	0.799	0.790
F1 macro	0.784	0.774
AUROC	0.944	0.944
F1 Bacteria	0.891	**0.902**
F1 Virus	0.680	0.689
F1 Normal	0.780	0.732

Table 7 - Results: Flat vs Hierarchical (DenseNet121) mean ± std over 3 seeds

Ensemble & Calibration Analysis

Calibration (ECE)

Transfer models are over-confident (ECE 0.15–0.16); the custom CNN is best- calibrated (0.11). 3-class single models: ECE ≈ 0.13–0.15.

3-class ensemble: ECE ≈ 0.05 — averaging 6 checkpoints smooths over-confidence, best-calibrated of all models.

Ensemble (6 checkpoints)

- ▶ Averages probabilities of ResNet18 + DenseNet121 × 3 seeds
- ▶ Binary: F1 = 0.896, recall = 0.997 (1 miss / 390)
- ▶ 3-class: best on every macro metric
- ▶ Reduces variance → more robust than any single model

Decision Thresholds (Binary)

No retraining needed → just move the threshold.

Two clinical operating points from the same model:

- ▶ Screening: max recall (≥ 99%)
- ▶ Assisted dx: max F1 (balanced)

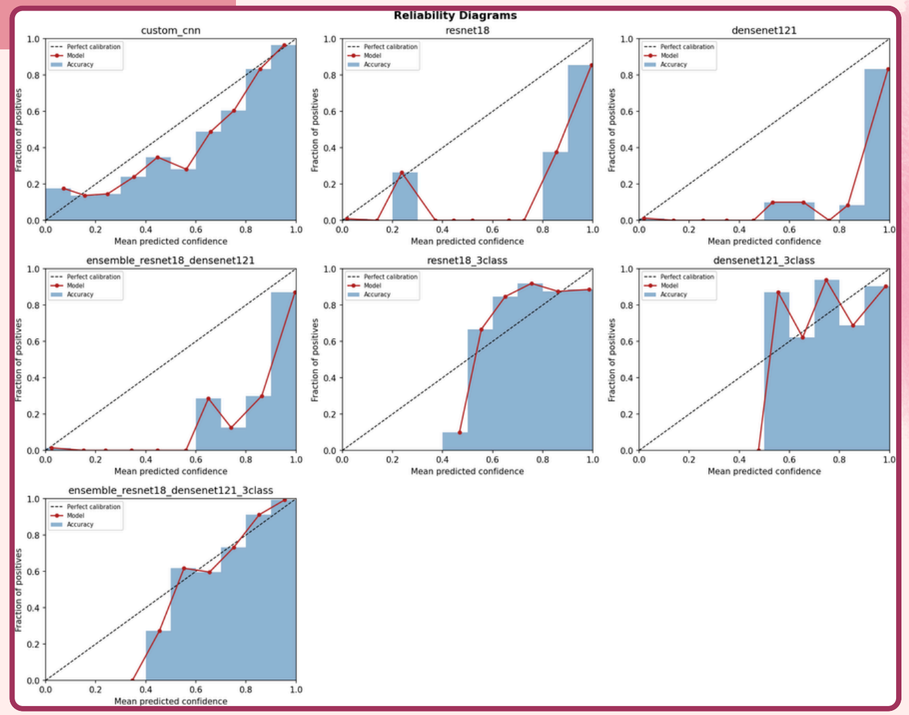


Fig. 9 - Reliability diagrams: predicted confidence vs. accuracy

GradCam: Explainability

What is Grad-CAM?

▶ Gradient of the predicted class score w.r.t. last conv layer
▶ Produces a heatmap of discriminative regions
▶ No architectural changes needed

Why It Matters Here

▶ Verify model attends to lung regions not artefacts
▶ Detect spurious correlations (text, borders)
▶ Build clinical trust in predictions

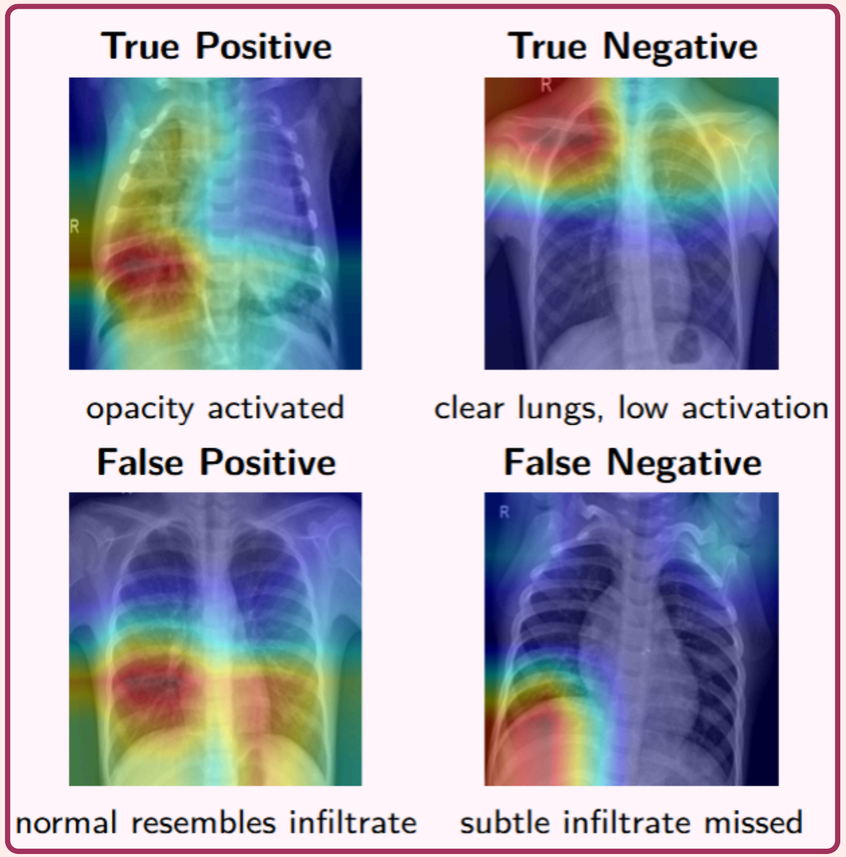


Fig. 10 - Grad-CAM heatmaps for DenseNet121: four binary outcome types (TP, TN, FP, FN)

GradCam: ResNet18 vs. DenseNet121

Interpretation

Both models attend to clinically plausible lung regions → not to text labels, borders, or artefacts.

DenseNet121 produces more focused, spatially coherent activations, consistent with its higher AUROC.

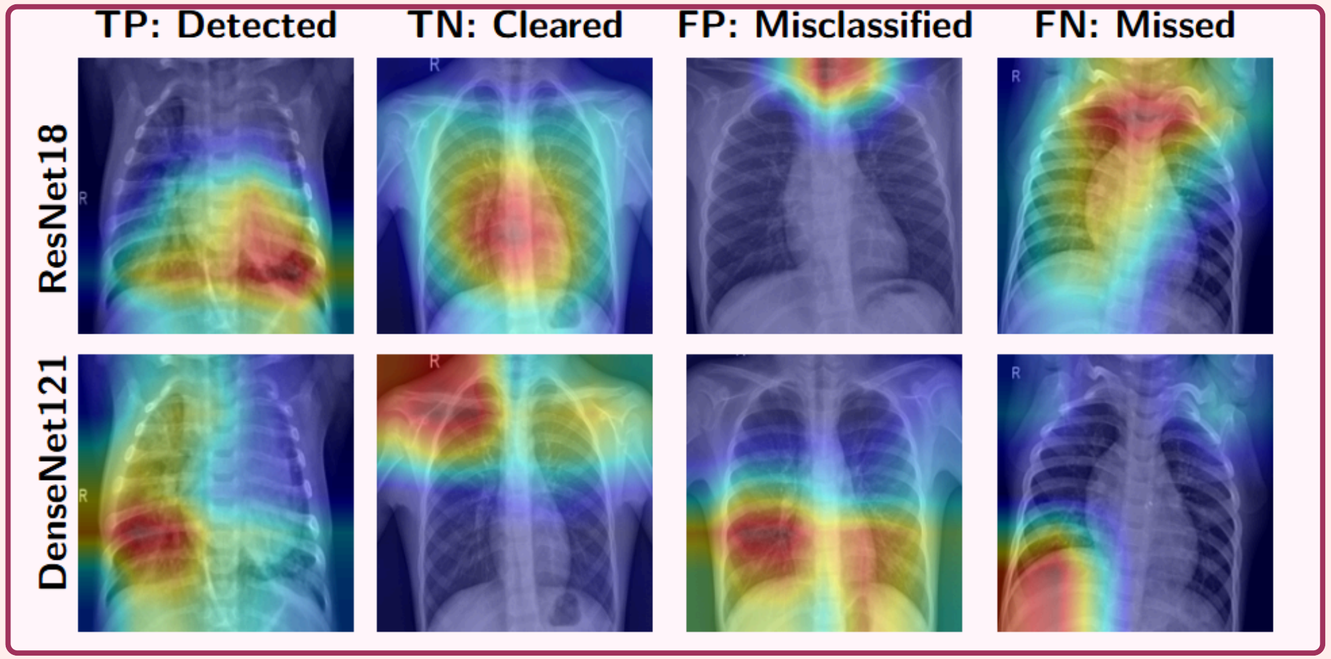


Fig. 11 - Grad-CAM comparison: ResNet18 vs. DenseNet121 across the four binary outcome types (TP, TN, FP, FN)

GradCam: Three-class Model

Interpretation

Correct bacterial/viral predictions focus on the relevant lung fields. The Normal → Virus failure shows diffuse activation over otherwise normal lungs → visually explaining the dominant error mode of the 3-class task.

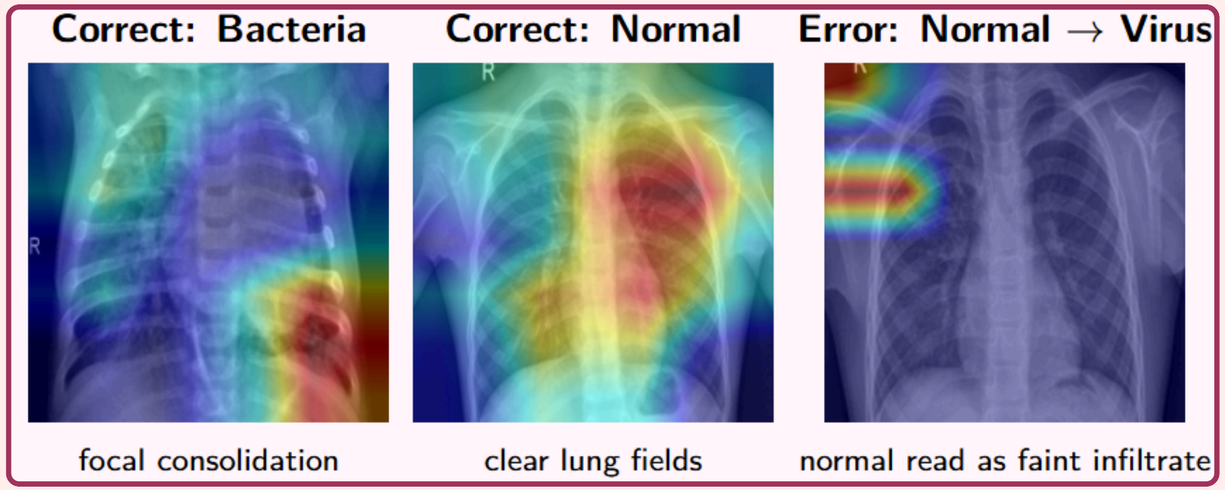


Fig. 12 - Grad-CAM heatmaps for DenseNet121 (3-class): correct Bacteria, correct Normal, and dominant error mode Normal → Virus

Ethics & Limitations

Potential Biases

- Single institution: Guangzhou → limited generalisability
- Pediatric only: may not transfer to adults or other equipment
- Class imbalance ⇒ tendency to over-predict pneumonia
- JPEG artefacts / text markers may be spuriously used

Technical Limitations

- 3-class labels inherit dataset's original labelling
- Validation split: only 16 images
- No external validation cohort
- Grad-CAM is an explanation, not ground-truth

Responsible Use Statement: NOT suitable for clinical deployment, educational prototype only.
Real deployment requires:

- Multi-site external validation
- Calibration & bias audit
- Clinician oversight (human-in-the-loop)
- Regulatory & ethical approval

Conclusion

Key Findings

▶ Binary ensemble: F1 = 0.896, recall = 0.997 (only 1 miss / 390), AUROC = 0.963
▶ 3-class innovation: macro-F1 = 0.808, AUROC = 0.960 (leakage-free patient-aware split)
▶ Hierarchical (Stage A: binary + Stage B: bacteria/virus) does not beat flat softmax (F1 0.774 vs 0.784) — VIRUS is a data-level bottleneck, not a modelling issue
▶ Transfer learning ≫ training from scratch; ensembling adds robustness
▶ Calibration & thresholds give 2 clinical operating points
▶ Grad-CAM confirms clinically plausible focus

Future Work

▶ Class imbalance: weighted / focal loss
▶ Temperature scaling for calibration
▶ Multi-institution / adult datasets
▶ Severity grading / segmentation
▶ External validation cohort
▶ Prospective clinical evaluation

References

Papers & Methods

1. Kermany et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 172(5):1122–1131. doi.org/10.1016/j.cell.2018.02.010

2. Rajpurkar et al. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225. arxiv.org/abs/1711.05225

3. He et al. (2016). Deep Residual Learning for Image Recognition. CVPR, 770–778. doi.org/10.1109/CVPR.2016.90

4. Huang et al. (2017). Densely Connected Convolutional Networks. CVPR, 4700–4708. doi.org/10.1109/CVPR.2017.243

References

Papers & Methods

5. Selvaraju et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. ICCV, 618–626. doi.org/10.1109/ICCV.2017.74

6. Krizhevsky et al. (2012). ImageNet classification with deep convolutional neural networks. NeurIPS, 1097–1105. doi.org/10.1145/3065386

7. Deng et al. (2009). ImageNet: A large-scale hierarchical image database. CVPR, 248–255. doi.org/10.1109/CVPR.2009.5206848

References

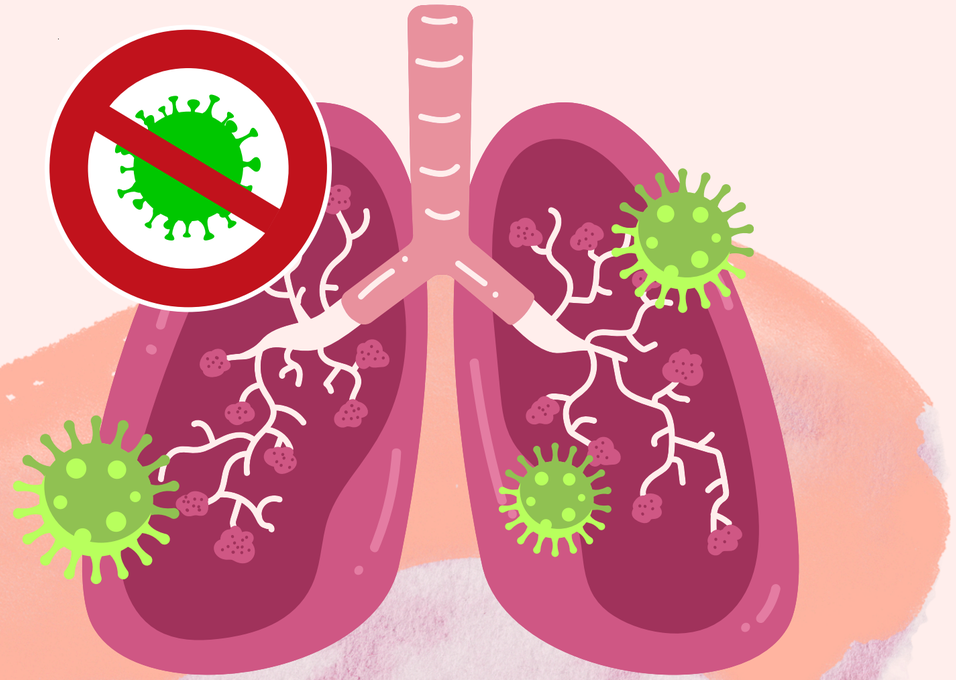
Dataset

Mooney, P. (2018). Chest X-Ray Images (Pneumonia). Kaggle (data from Kermany et al., 2018). kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

Learning Resources

Vizuara Newsletter:

▶ ResNet: The architecture that changed ML forever.
vizuaranewsletter.com/p/resnet-the-architecture-that-changed
▶ DenseNet and EfficientNet are 1/20th the size of VGG16...
vizuaranewsletter.com/p/densenet-and-efficientnet-are-120th



THANK YOU

luanacarolina@ua.pt · jakub.blaszczyk@ua.pt
University of Aveiro, DETI · May 2026