Thoracic Findings Classifier (DenseNet121)

From preprocessing → modeling → evaluation (executive summary)

Best checkpoint (high-res pass)

Macro AUROC (test): 0.741 | Test AUC (Keras): 0.740 | Restored best epoch: 3

Key bottlenecks: Pneumonia + Nodule (lowest AUROC). Next levers: class-weighted BCE + threshold tuning.

Multi-label | AdamW | Early stopping

Project synopsis

Why this project

Rapid triage support for thoracic findings in chest X-rays.
Multi-label setup: each study can have multiple pathologies.
Primary KPI: AUROC (macro + per-label) to handle imbalance.

What was built

DenseNet121 backbone + multi-label classification head.
Training in phases: frozen backbone → fine-tune → optional 320px pass.
Evaluation outputs: macro AUROC, per-label AUROC, CSV-ready metrics.
Key result: macro AUROC (test) ≈ 0.741 with stable generalization.

Scope: preprocessing → modeling (this deck).

Data & problem framing

Task definition

Input: single chest X-ray image.
Output: 8 independent probabilities (sigmoid) — one per label.
Loss: binary cross-entropy across labels.

Core constraints

Label imbalance (rare positives) + label noise (esp. Pneumonia).
Small findings (Nodule) benefit from higher resolution.
Primary selection metric: AUROC (macro + per-label).

Target labels

1. Atelectasis
2. Cardiomegaly
3. Effusion
4. Infiltration
5. Mass
6. Nodule
7. Pneumonia
8. Pneumothorax

Multi-label thoracic pathology classification.

Preprocessing pipeline (image → tensor)

Ingest & validate

Resolve paths
Drop corrupt images

Label curation

Multi-label vector
Split by patient

Leakage controls

Split before augmentation.
Patient-level split if available.
No val/test transforms beyond resize/normalize.

Resize

224px base
320px optional pass

Normalize

Scale to [0,1]
Mean/Std if used

Augment (train only)

Flip, mild rotate
Small zoom/contrast

tf.data performance

cache/prefetch
parallel map

Reproducibility

Fixed seeds.
Deterministic preprocessing functions.
Versioned checkpoints + CSV metrics exports.

Preprocessing is lightweight and model-compatible (DenseNet).

Modeling (DenseNet121 multi-label head)

Backbone

DenseNet121 pretrained on ImageNet (transfer learning).
Global Average Pooling to compress feature maps.
BatchNorm layers typically frozen during fine-tuning for stability.

Classification head

Dense(8) + sigmoid → per-label probability
Loss: Binary Cross-Entropy (BCE).
Optional: class-weighted BCE to address rare positives.

Training configuration

Optimizer: AdamW (weight decay = 1e-4).
Metric: multi-label AUROC (macro + per-label reporting).
Callbacks: EarlyStopping + ReduceLROnPlateau + ModelCheckpoint.
Stages:

1) Frozen base (stabilize head)
2) Fine-tune top layers (unfreeze subset)
3) Optional high-res (320px) pass for small findings

DenseNet is fast, strong for medical imaging baselines.

Fine-tuning strategy (what changed, and why)

Training phases

Phase 1: Frozen base

Train head + stabilize

Phase 2: Fine-tune

Unfreeze last N layers

Phase 3: 320px pass

Best epoch restored: 3

Why high-res can help

Some pathologies are small or subtle (e.g., Nodules).
A short 320px pass can preserve fine detail.
Keep it controlled: unfreeze fewer layers + freeze BatchNorm.

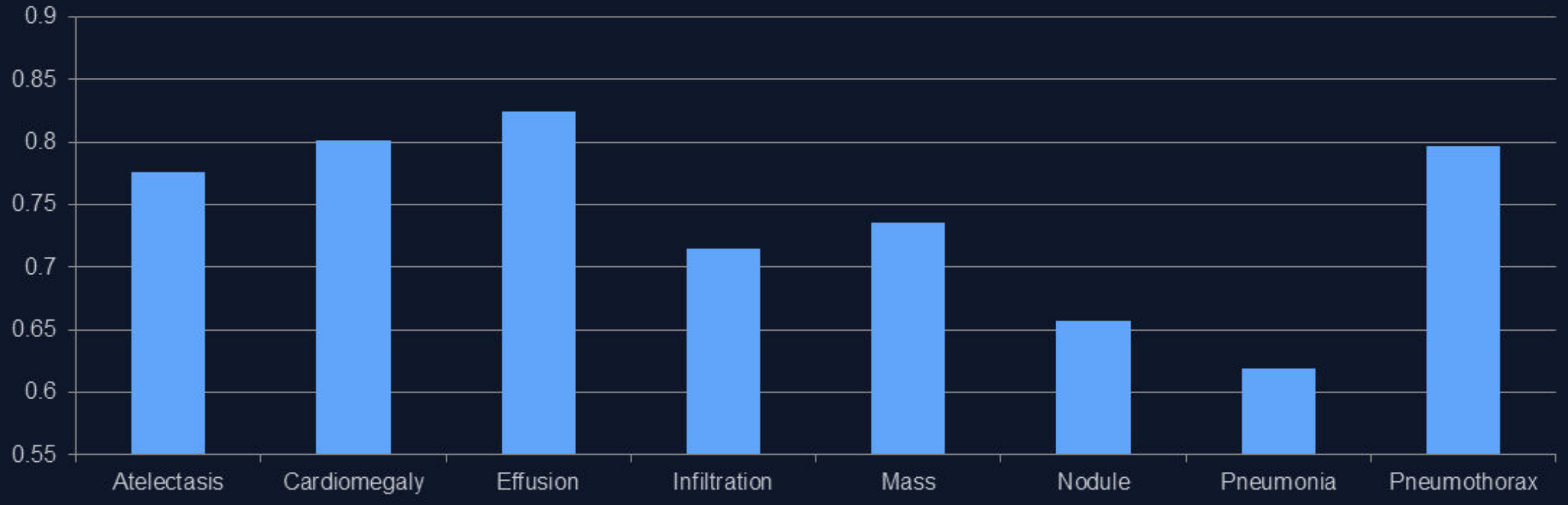
Why it can slow down

320px increases compute per step (more pixels).
Unfreezing too many layers increases backprop cost.
Multi-label AUC is heavy; consider val-only metric computation.

Best practice: small unfreeze (25–50 layers) + low LR (≈1e-5) for 1–3 epochs at 320px.

Evaluation (test): macro + per-label AUROC

Macro AUROC: 0.741 (Val AUC: 0.767 | Test AUC: 0.740)



Primary bottlenecks

Pneumonia (0.619) and Nodule (0.657)

Recommended levers

Weighted BCE (pos_weight per label) + per-label threshold tuning for F1.

AUROC computed per label (test). Update label names if your dataset order differs.

Action plan to lift macro AUROC & F1

1) Class imbalance mitigation

Implement per-label pos_weight in BCE (cap extreme weights).
Optionally oversample rare positives for Pneumonia/Nodule.
Track per-label AUROC deltas, not just macro.

2) Threshold tuning (for F1)

Tune thresholds on validation per label (0.05–0.95 grid).
Report macro/micro/weighted F1 on test.
Use fixed thresholds for final model packaging.

3) Controlled 320px pass (only if needed)

Goal: help small findings without overfitting.
Settings:

- IMG_SIZE = 320; batch reduced as needed
- Unfreeze last 25–50 layers (not 200)
- Freeze BatchNorm layers
- LR ≈ 1e-5 (or 5e-6) for 1–3 epochs

Success criteria:

- Nodule/Pneumonia AUROC improves without harming others
- Macro AUROC increases on test

Make changes one at a time; compare per-label deltas to baseline.

Artifacts produced (reproducible handoff)

What to ship

- Best checkpoint file (ModelCheckpoint output).
- Config snapshot: IMG_SIZE, batch size, LR schedule, N_UNFREEZE.
- Evaluation CSVs:
  - eval_summary.csv (macro + val/test AUC)
  - eval_per_label_auroc.csv (per label)
  - confusion_template.csv (TP/FP/TN/FN + thresholds)
- Notebook(s): preprocessing + training + evaluation.

How to reproduce

1) Build datasets (train/val/test) with identical preprocessing.
2) Train frozen-base stage; save checkpoint.
3) Fine-tune (unfreeze subset + BN frozen).
4) Run evaluation script to export CSV metrics.
5) Tune thresholds (val) and re-evaluate on test.

Deliverables focus on auditability and repeatability.

Executive close

Current performance (test)

Macro AUROC: 0.741
Best labels: Effusion (0.825), Cardiomegaly (0.801)
Weak labels: Pneumonia (0.619), Nodule (0.657)

Key message

The pipeline is stable and reproducible. Performance gains are most likely from handling rare-label imbalance and optimizing thresholds—especially for Pneumonia and Nodule.

End.