Capstone 3 Project Report
Multi-Label Thoracic Finding Classification with DenseNet121

Author: Albert Vempala

Focus: why, how (preprocessing + modeling), evaluation, and business impact. Metrics and figures in this report reflect the values shared during experimentation in this chat.

Headline metric (test, sklearn):	Macro AUROC = 0.7406
Keras AUC (validation):	0.7666
Keras AUC (test):	0.7400
Best epoch restored (high-res pass):	3
High-res input size:	320px

Executive summary

A DenseNet121 transfer-learning model was trained for multi-label prediction of eight thoracic findings. On the held-out test set, the project achieved a **macro AUROC of 0.7406** (sklearn). A high-resolution pass at 320px used early stopping and restored weights from the best epoch (3) to manage overfitting.

Performance varies by finding: Effusion, Cardiomegaly, and Pneumothorax show strong discrimination, while Pneumonia and Nodule remain the bottlenecks. Subsequent work should focus on those weak labels via imbalance-aware training and threshold tuning.

1. Why

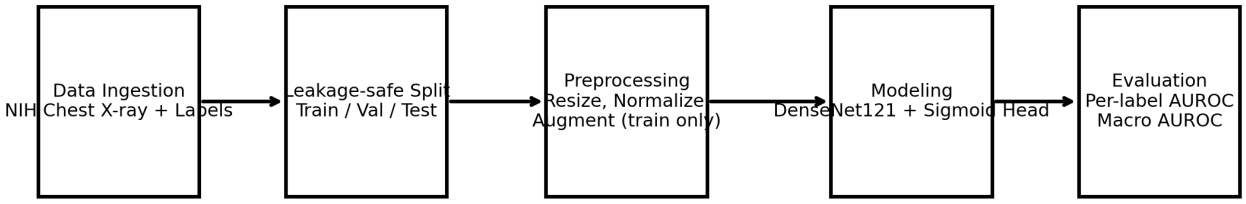
Chest X-ray interpretation is a high-volume workflow where prioritization and decision support can improve operational efficiency. A multi-label classifier can generate per-finding risk scores that help route studies for review, support auditing, and enable aggregate analytics. This capstone focuses on building a reproducible pipeline and quantifying discrimination performance using AUROC.

This project is not a clinically validated diagnostic system. Any real-world use would require external validation, calibration, and governance.

2. How

2.1 Pipeline overview

The pipeline is designed to be leakage-safe and reproducible: data ingestion, train/validation/test split, preprocessing and augmentation, model training with transfer learning, and evaluation on held-out data.



2.2 Preprocessing

Preprocessing focuses on consistent transforms across splits: deterministic resizing to the chosen input resolution, normalization consistent with the pretrained backbone, and training-only augmentations for robustness. Evaluation splits (validation/test) use deterministic preprocessing only.

3. Modeling

3.1 Architecture

The model uses a pretrained DenseNet121 backbone with a sigmoid multi-label head. This enables transfer learning: reuse general visual features while training a task-specific classifier for thoracic labels.

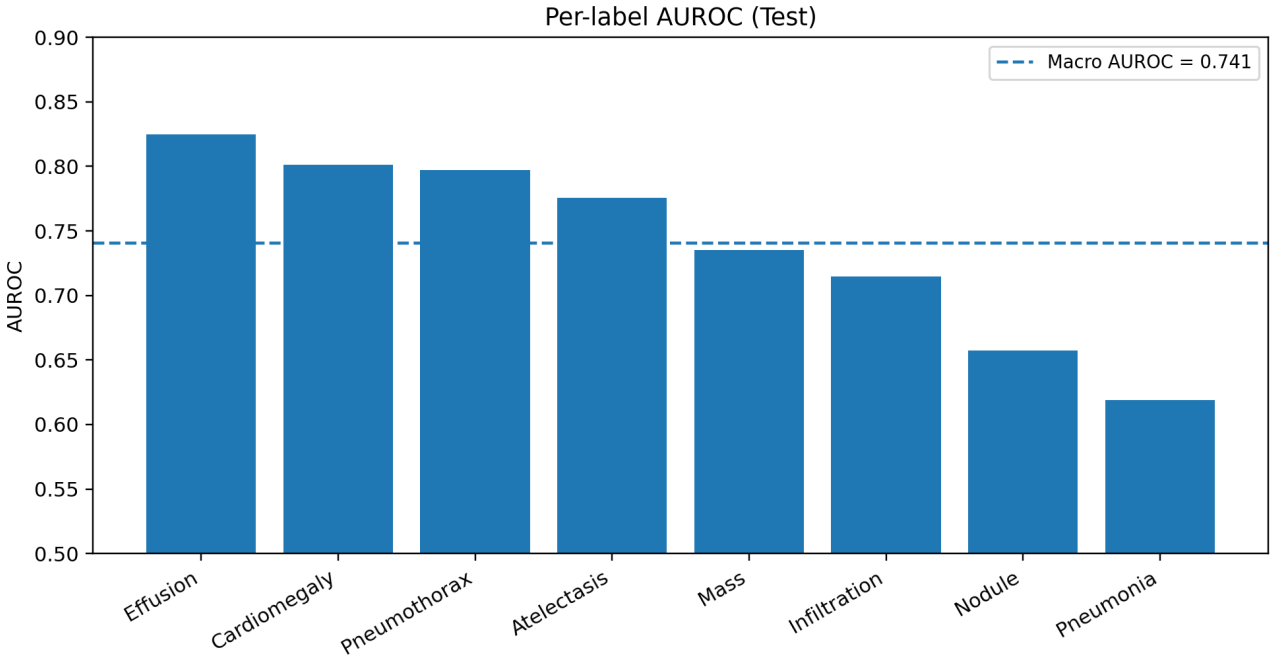
3.2 Training strategy

Training follows a staged approach: start with the backbone frozen to stabilize the head, then fine-tune a subset of layers. For the high-resolution stage, the input was set to 320px and early stopping restored weights from epoch 3.

Observed AUC values during evaluation were 0.7666 on validation (Keras AUC) and 0.7400 on test (Keras AUC). Macro AUROC (sklearn) is reported separately as the average of the per-label AUROCs below.

4. Evaluation

For multi-label problems, AUROC is computed per label (one-vs-rest). Macro AUROC averages AUROC across labels, weighting each label equally. This is useful when labels differ in prevalence.



4.1 Per-label AUROC (test)

Label	AUROC (test)
Effusion	0.8246
Cardiomegaly	0.8012
Pneumothorax	0.7970
Atelectasis	0.7756
Mass	0.7353
Infiltration	0.7148
Nodule	0.6575
Pneumonia	0.6187

Headline metric: Macro AUROC (test) = **0.7406**. The lowest-performing labels (Pneumonia, Nodule) are the most efficient improvement targets for raising the macro score.

5. Business impact

Potential value (after proper validation) comes from using predicted risk scores as workflow signals:

• **Queue prioritization:** route studies for review based on per-finding risk scores.
• **Second-signal review:** highlight studies with elevated predicted risk to reduce misses when used alongside human interpretation.
• **Operational analytics:** aggregate label scores to monitor trends and support quality programs.

Given macro AUROC ~0.74 across eight findings, the most defensible operational use is decision support with human oversight, plus monitoring and periodic recalibration.

6. Recommendations

Next steps should focus on improving weak labels and strengthening evaluation artifacts:

• **Imbalance-aware training:** apply per-label positive-class weighting (pos_weight) or controlled sampling.
• **Threshold tuning:** select per-label thresholds on validation to optimize F1/precision-recall trade-offs.
• **Controlled high-res fine-tune:** fine-tune only a small number of top layers (e.g., N_UNFREEZE ~ 25) with BatchNorm frozen and low LR.
• **Reporting:** export y_true and y_score to generate true ROC and PR curves for GitHub-ready figures.

7. Limitations

• This report summarizes AUROC values shared during experimentation; true ROC curve plots require the underlying FPR/TPR points derived from (y_true, y_score).
• Label uncertainty and prevalence differences can materially affect per-label performance.
• Any clinical use requires external validation, calibration, and governance.