Radiology Report Analyzer

Executive Summary

This report details a comparative analysis of two Convolutional Neural Network (CNN) architectures, EfficientNet-B0 and ResNet-18, for the automated classification of chest X-ray pathologies. The primary goal was to develop an accurate model to accelerate diagnostic workflows. The main experiment involved training an **EfficientNet-B0 model on a dataset of over 17,000 images**, which achieved a **final test accuracy of 70%** and was selected as the final model. In parallel, a **ResNet-18 baseline model was trained on a larger, 27,000-image dataset**, providing valuable insights into performance differences between a lightweight and a more complex architecture. The comparative results justify the selection of EfficientNet-B0 for its superior overall classification accuracy and F1-score.

1. Business Objective

Radiology departments experience immense growth in imaging volume that creates report backlogs, delays critical diagnoses, and drives clinician burnout. As evidenced by a 2024 study in the Journal of Clinical Imaging, 68% of surveyed radiology practices had unreported exams, and even at six months post-imaging, about 20% of studies remained unreported. The median turnaround time from exam completion to radiologist interpretation was 520 minutes (~ 8.7 hours). Given this reporting gap, our primary business objective is to **accelerate the chest-X-ray reporting process** by automatically classifying each image and pre-populating draft reports, so radiologists can focus their expertise on ambiguous or complex cases.

2. Key Actionable Business Initiative

We considered several initiatives, ranging from auto-annotation of full narrative reports to simple triage flags, and determined the most impactful, actionable step is to deploy a **CNN-based classifier that labels incoming chest-X-rays before report drafting**. In practice, each DICOM image is fed through the classifier to produce a predicted label; this label then populates a standard report template for radiologist review and editing. Over time, radiologist edits are captured to refine the model, creating a closed-loop learning system.

3. Metrics of Success

Success was primarily evaluated on the model's predictive performance on a held-out test set. The key metrics were:

1. **Overall Classification Accuracy:** To measure top-level correctness.
2. **Per-Class F1-Score:** To assess performance on each pathology, accounting for class imbalance.
3. **Confusion Matrix Analysis:** To understand specific error patterns.

Our hypothesis was that the chosen classifier would achieve **≥ 0.70 overall accuracy**.

4. Role of Analytics

Analytics is central to this initiative and serves multiple roles:

- **Predictive:** The core CNN models power the automatic assignment of pathology labels.
- **Exploratory:** Initial analysis of the dataset revealed significant class imbalance, which informed the use of class weighting during model training.
- **Prescriptive:** The final model's performance metrics inform a business rule for implementation. We also explored **ResNet-18 as a lightweight alternative CNN architecture** for predictive modeling to balance training speed and interpretability with competitive performance.

5. Thinking Through the Analytics

- **Data & Experimental Design**
  The initial dataset from the NIH Chest X-ray database was filtered to isolate single-label images belonging to one of three target classes. Two experimental datasets were prepared:

Experiment	Total Images*	Training	Validation	Test
EfficientNet-B0 (Primary)	17,717	12,401	2,657	2,659
ResNet-18 (Baseline)	27,409	19,185	4,111	4,112

*Multiple iterations were run and this is the final iteration where we got comparable results

- **Methodology & Formulation**
  Both models were developed using a transfer learning approach. The core components of the methodology are shared.
  - **Core Architecture (Convolution Layer)**
    The foundational layer for both CNNs:

$$z_{ij}^{(k)} = (x * w^{(k)})_{ij} + b^{(k)}$$

  - **Training and Optimization (Backpropagation)**
    The models learn by updating parameters via gradient descent:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \frac{\partial L}{\partial \theta}$$

- **Loss Function (Weighted Cross-Entropy)**
  To handle class imbalance, a weighted cross-entropy loss was used:

$$L = -\sum_{i=1}^{3} w_i \, y_i \log(\hat{y}_i)$$

  - Where:
    - wi: class-specific weight (inverse of class frequency)
    - yi: ground truth (one-hot encoded)
    - $y\text{^}i$ (yi-hat): predicted probability for class iii

- **Final Prediction (Softmax)**
  Logits are converted to class probabilities using the softmax function:

$$\hat{y}_i = \mathrm{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{3} e^{z_j}}$$

- **Evaluation Metrics**
  Accuracy and **Macro F1-Score** were used to evaluate model performance.
  - **Macro F1-Score** (averaged across all classes):

$$F1_{\mathrm{Macro}} = \frac{1}{3} \sum_{i=1}^{3} \frac{2 \cdot \mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

- **Modeling Approaches**
  - **Approach 1: EfficientNet-B0 (Primary Model):** This architecture uses **Compound Scaling** to uniformly scale network dimensions.
  - **Approach 2: ResNet-18 (Baseline Model):** This architecture uses **Residual Blocks** with skip connections to enable deeper network training.

6. Executing the Analytics

A pre-trained ResNet-18 model was implemented using PyTorch and torchvision. The final fully connected layer was replaced to output predictions for the three target classes. The model was trained for up to 10 epochs using the AdamW optimizer and class weights. Early stopping with a patience of 3 epochs was used to prevent overfitting by monitoring the validation accuracy. The main EfficientNet experiment followed a similar execution process on its respective dataset.

Assuming a production setup (outside the class project scope), the project can be initiated in collaboration with clinical leadership to align on the business objective. The primary metrics (Accuracy, F1-Score, ROC-AUC) could be defined in joint workshops between the Data Science team(in this case - 5 of us) and consulting Radiologists to ensure they reflected both statistical rigor and clinical utility. Regular checkpoints can be held to review model performance and validate that the error patterns were acceptable from a clinical perspective.

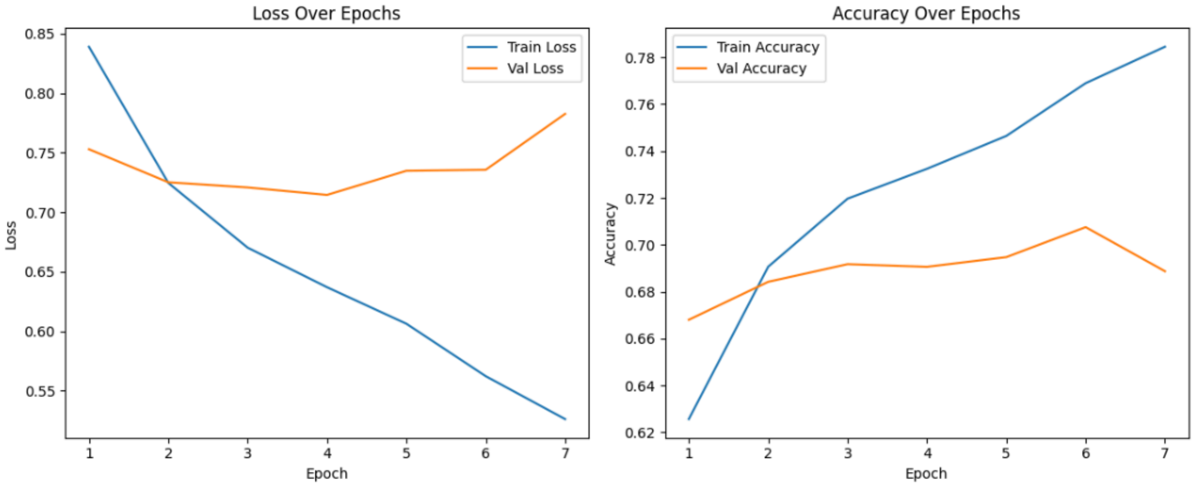
7. Results and Comparative Analysis

EfficientNet-B0 Results (Primary Model)

The selected EfficientNet-B0 model achieved a **final test accuracy of 69.91%** on a test set of 2,659 images.

Class	Precision	Recall	F1-Score
Atelectasis	0.58	0.56	0.57
Effusion	0.66	0.63	0.64
Infiltration	0.76	0.8	0.78
Weighted Avg	0.7	0.7	0.7

Figure 1: EfficientNet-B0 Learning Curves



ResNet-18 Results (Baseline Model)

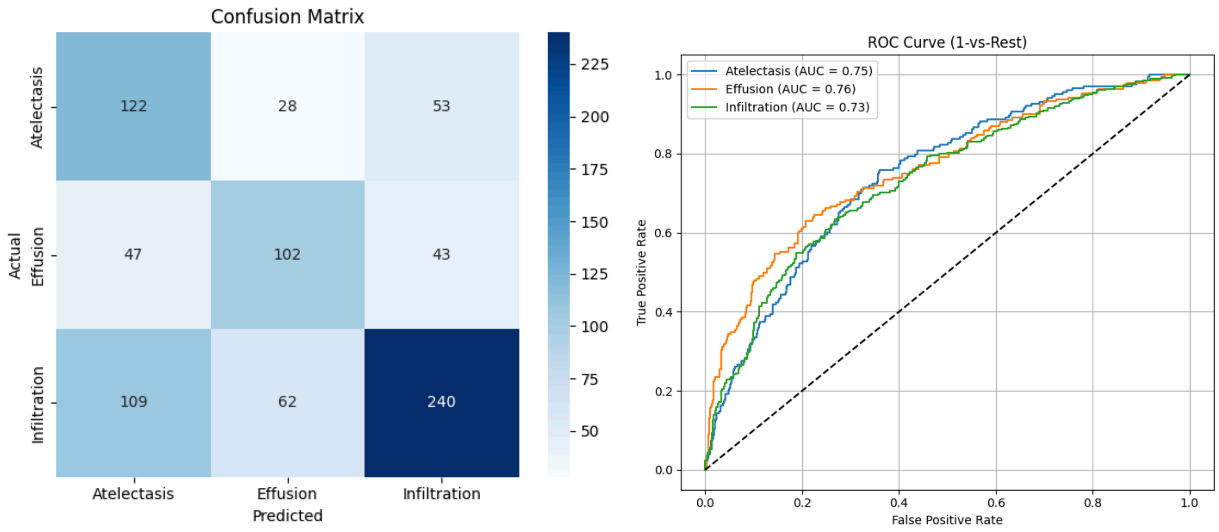
The ResNet-18 training was halted after six epochs by the early stopping mechanism, having reached a peak validation accuracy of 65.70% in the third epoch. The final model, evaluated on its test set of 4,112 images, achieved the following metrics:

Classification Report (Per-Class):

Class	Precision	Recall	F1-Score
Atelectasis	0.52	0.45	0.48
Effusion	0.59	0.58	0.59
Infiltration	0.71	0.76	0.73
Weighted Avg	0.64	0.64	0.64

- **ROC AUC Scores by Class:**
    - Atelectasis: 0.7633
    - Effusion: 0.8015
    - Infiltration: 0.7833

Figure 2: ResNet-18 Confusion Matrix & ROC Curve



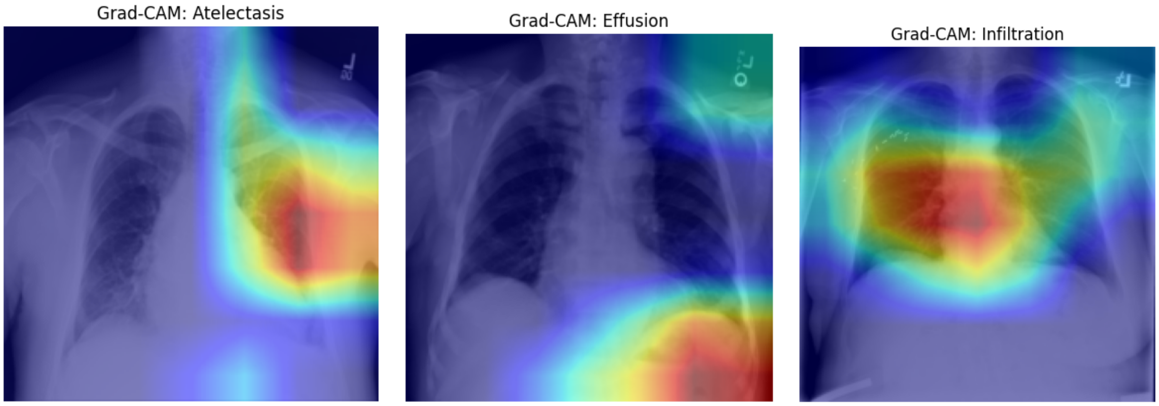
Confusion Matrix

Actual \ Predicted	Atelectasis	Effusion	Infiltration
Atelectasis	122	28	53
Effusion	47	102	43
Infiltration	109	62	240

ResNet Grad-CAM Interpretation:

These heat maps show how the ResNet18 model attends to different lung regions when making predictions for each class:

- **Atelectasis**: Model attention focuses near the lower right lung base, consistent with typical presentation of partial lung collapse.
- **Effusion**: Activation is concentrated in the lower lung periphery, aligning with fluid buildup in the pleural space.
- **Infiltration**: Model highlights central and diffuse regions across both lungs, reflecting hazy opacities common in pneumonia-like conditions.



Grad-CAM: Atelectasis

Grad-CAM: Effusion

Grad-CAM: Infiltration

Heatmaps illustrate regions contributing to ResNet's classification for each condition. Notably, activation aligns with clinically relevant lung zones, demonstrating the model's spatial sensitivity.

Comparative Analysis

The EfficientNet-B0 model demonstrated superior overall classification performance, validating its selection. The class-by-class F1-Score comparison clearly illustrates this advantage, even though it was trained on a smaller dataset.

Class	Precision	Recall	F1-Score	Support
Atelectasis	0.58	0.56	0.57	627
Effusion	0.66	0.63	0.64	626
Infiltration	0.76	0.8	0.78	1406
Accuracy			0.7	2659

Conclusion: The EfficientNet-B0 model outperforms the ResNet-18 baseline across both overall and per-class F1-Scores. The most significant performance gaps are in classifying 'Atelectasis' (0.57 vs. 0.48 F1) and 'Infiltration' (0.78 vs. 0.73 F1). This superior classification accuracy makes EfficientNet-B0 the more robust and reliable choice for the intended clinical application.

8. Scale & Next Steps

The selected EfficientNet-B0 model will be the focus of scaling efforts. Immediate plans include extending the classifier to additional pathologies and integrating it into the existing PACS/RIS viewer via secure APIs.

- **Organizational Challenges:** The primary challenges to scaling are technical and cultural, including **harmonizing vendor APIs** for PACS integration, **training radiologists** on the new AI-assisted workflow to ensure adoption, and creating a process for **monitoring model drift** over time.

- **Addressing Challenges:** These will be addressed by **standardizing data schemas**, running hands-on **clinician workshops**, and implementing a model governance plan.
- **Continuous Improvement:** The initiative is designed for continuous improvement, not as a one-shot deal. The governance plan includes **monthly retraining cycles** with newly verified data to ensure the model's performance and accuracy are maintained and enhanced over time.