

# Lossless Mechanistic Compression and Surgical Correction of Medical Imaging Models

Yeonseong Cynn<sup>1</sup>

<sup>1</sup>River Lab

May 2026

## Abstract

Medical imaging models such as CheXNet (DenseNet121) are widely deployed for multi-label thoracic pathology classification but suffer from large parameter counts (6.97M) and opaque debugging pipelines. We present a unified framework that (1) **losslessly compresses** the model by 51.43% (6.97M  $\rightarrow$  3.38M parameters) via channel-wise sparsity-constrained weight reconstruction on NIH ChestX-ray14, with AUROC change of +0.0004 and per-image latency reduced from 15.17 to 14.73 ms; (2) enables **surgical correction** through classifier-channel attribution and selective weight zeroing (5-channel correction reduces a target false positive by  $\Delta\text{prob} -0.13$  with zero true-positive loss and exactly zero AUROC change on the other 13 pathologies); (3) provides a **cost-aware Treatment Decision System** routing each pathology issue to its cheapest effective intervention; (4) supports **clinical report auto-generation** with channel-level evidence, Grad-CAM region mapping, and mutual-exclusivity-based exclusion. We further discover that polarized classifier channels are not architectural conflicts but *bipolar discriminative axes* exploiting label mutual exclusivity (Jaccard  $< 0.1$  in 89 of 100 polarized channels). On CheXNet, per-class Youden’s- $J$  threshold calibration alone (no retraining) raises the cohort-average  $F_1$  by  $\times 1.6$  and recall by  $\times 7$  relative to the default threshold-0.5 operating point; classifier-head fine-tuning (18K trainable parameters,  $\approx 85$  s) adds at most marginal additional  $F_1$  and regresses several best-performing classes, so we characterize threshold calibration as the operationally meaningful intervention and fine-tuning as a mean-AUROC stabilizer. Compression method specifics are proprietary; model weights, inference code, and analysis scripts are released for reproduction.

## 1 Introduction

Multi-label thoracic pathology classifiers such as CheXNet [10], built on DenseNet121 [3], are widely deployed for chest X-ray analysis but face three orthogonal operational challenges. First, their parameter footprint (6.97M for the open CheXNet weights) limits edge deployment in hospitals. Second, the model is typically treated as a black box for clinicians; standard saliency methods such as Grad-CAM [11] answer “where the model looked” but not “what features were used” or “how to fix specific errors.” Third, debugging operational issues (false positives, missed detections, distribution shifts to new hospitals) tends to default to expensive full retraining, even when a cheaper intervention suffices.

We address all three challenges in a single mechanistic framework. Our starting point is a lossless compression of CheXNet that preserves the model’s output to numerical precision while shrinking it by 51.43%. The resulting compressed model has the additional property that the number of effectively contributing channels per layer is small, which makes channel-level attribution and direct intervention (*surgical correction*) practical—an option not available in the dense original. We use

this property to (i) discover that the “polysemantic” channels of the compressed classifier are not conflicting but rather encode mutual-exclusivity binary axes between pathology pairs that rarely co-occur in the data; (ii) introduce *surgical correction*—a selective weight zeroing on the classifier head that removes a target false positive without affecting other pathologies; (iii) develop a Treatment Decision System that, given diagnostic signals about a pathology’s failure mode, routes the user to the cheapest effective fix; and (iv) auto-generate clinical reports that combine probabilities, Grad-CAM regions, channel-level reasoning, and mutual-exclusivity-based exclusion.

### Contributions.

1. Channel-wise sparsity-constrained weight reconstruction that compresses CheXNet by 51.43% with mean-AUROC change +0.0004 on a 1,045-image NIH test split, per-pathology max  $|\Delta\text{AUROC}| \leq 0.0033$  (within sampling noise), and output identity to numerical precision (max  $|\Delta\text{logit}| < 5 \times 10^{-6}$ ).
2. Discovery and quantification of *bipolar discriminative axes*: polarized classifier channels are mutual-exclusivity exploitations, not conflicts (Jaccard-based legitimacy: 89 of 100 polarized channels are perfect or legitimate, zero conflict).
3. *Surgical correction* for mechanistic intervention: 5-channel classifier weight zeroing softly reduces a target false positive’s probability by  $\Delta\text{prob} -0.13$  (a confidence shift; the threshold-0.5 decision boundary is not crossed at this  $K$ ) with zero true-positive loss and exactly zero AUROC change on the other 13 pathologies (row-independent classifier weights guarantee isolation by construction).
4. A cost-aware Treatment Decision System routing pathology issues to threshold calibration, surgical correction, partial retraining, or augmentation; empirical cost matrix updated from initial estimates.
5. Mechanistic clinical report auto-generation combining channel-level reasoning with conventional spatial explanations.

## 2 Background and Related Work

**Multi-label chest X-ray classification.** CheXNet [10] adapted DenseNet121 [3] for chest X-ray analysis. Subsequent multi-source training combining NIH ChestX-ray14 [13], CheXpert [4], MIMIC-CXR [6], and others produced the “densenet121-res224-all” weights distributed through the *torchxrayvision* library [1], which we adopt as baseline.

**Channel pruning.** Channel pruning [2, 7] removes feature maps to reduce computation but typically incurs an accuracy drop or requires post-pruning fine-tuning. Our compression is closer in spirit to weight-reconstruction approaches [2]: rather than scoring and removing the least important channels, we re-fit the per-layer weight matrices with an  $L_1$  penalty that drives many output channels to exactly zero post-ReLU, after which we structurally remove the corresponding rows and columns.

**Mechanistic interpretability.** Olah et al. [8, 9] introduced channel-level visualization and noted the phenomenon of *polysemanticity*: a single channel often responds to multiple unrelated concepts. We revisit polysemanticity in the multi-label medical setting and show that what appears polysemantic in our compressed classifier is in fact *mutual-exclusivity exploitation*: the model uses one channel as a bipolar axis between two pathologies that almost never co-occur.

**Spatial explainability.** Grad-CAM [11], Score-CAM [12], and Layer-CAM [5] produce class-conditional spatial heatmaps. These methods explain *where* the model attends but not *which internal features* drive a particular prediction. Our framework complements spatial explanations with channel-level attribution and mutual-exclusivity reasoning.

**Threshold calibration.** In multi-label classification with imbalanced positive rates, threshold 0.5 is rarely optimal. Youden’s J statistic [14] selects the threshold maximizing sensitivity + specificity  $-1$  on a calibration set. We demonstrate that calibration combined with minimal retraining yields  $F_1 \times 1.6$  and  $\text{recall} \times 7$  over the default threshold.

### 3 Method

We work with CheXNet (DenseNet121, growth rate  $k = 32$ , four dense blocks of  $\{6, 12, 24, 16\}$  layers). Each dense layer has the bottleneck-growth structure  $\text{BN} \rightarrow \text{ReLU} \rightarrow 1 \times 1 \text{ conv} \rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow 3 \times 3 \text{ conv}$ . The  $1 \times 1$  *bottleneck* conv reduces the concatenated input to a small intermediate dimension, and the  $3 \times 3$  *growth* conv produces  $k = 32$  new feature maps that are concatenated to the block state.

#### 3.1 Channel-Wise Sparsity-Constrained Compression

We compress CheXNet through a channel-wise sparsity-constrained weight-reconstruction procedure applied to each dense layer. The bottleneck and growth convolutions of every layer are jointly re-fit to reproduce the original per-layer output under an  $L_1$ -style regularizer. After fitting, weights below a small magnitude threshold are zeroed, and intermediate channels whose post-ReLU activation never exceeds the threshold are structurally removed along with the corresponding rows and columns in the surrounding weights. The intermediate batch normalization can then be folded into surrounding linear operations and replaced by an identity module, which yields a  $\approx 15\%$  inference-latency reduction without changing the output. The precise loss formulation, the channel-inactivity criterion, and the structural-reduction protocol are proprietary; released artifacts (compressed weights, inference code, and analysis scripts) are sufficient for downstream reproduction at the result level, including the empirical findings in Sections 3.2–3.6.

**Practical considerations.** We note that naive implementations of this style of compression—uncritical  $L_1$  pruning followed by structural removal—fail to preserve output identity due to subtle interactions among the bottleneck refit, the channel-inactivity criterion, and module-state semantics during structural reduction. The released inference code reflects the validated configuration; specific implementation choices that resolve these interactions are proprietary procedural elements of the compression and are not disclosed here.

### 3.2 Polarized Channel Legitimacy Analysis

After compression we observe that many channels of the classifier head  $W \in \mathbb{R}^{P \times 1024}$  ( $P = 18$  pathologies) have both strongly positive and strongly negative weights across different output rows. We call such channels *polarized*:

$$c \text{ is polarized if } \exists p^+, p^- \text{ s.t. } W_{p^+,c} > \tau \text{ and } W_{p^-,c} < -\tau, \quad (1)$$

with  $\tau = 0.3$  in our experiments.

A naive interpretation would call this an architectural conflict: the same channel pushes one pathology up and another down. We argue that this is in fact the correct, efficient encoding when the two pathologies are *mutually exclusive* in the data. For each polarized pair  $(p^+, p^-)$  we compute the empirical Jaccard similarity

$$J(p^+, p^-) = \frac{|p^+ \cap p^-|}{|p^+ \cup p^-|} = \frac{\#\{i : y_{i,p^+} = 1 \wedge y_{i,p^-} = 1\}}{\#\{i : y_{i,p^+} = 1 \vee y_{i,p^-} = 1\}}. \quad (2)$$

We classify each polarized channel into one of four legitimacy categories based on its worst-case pair Jaccard: *perfect* ( $J = 0$ ), *legitimate* ( $J < 0.1$ ), *mixed* ( $0.1 \leq J < 0.3$ ), or *conflict* ( $J \geq 0.3$ ). Only *conflict* channels are evidence that the model is forced to compromise between two co-occurring pathologies.

### 3.3 Surgical Weight Correction

When a specific case is misclassified as positive for pathology  $p$ , we locate FP-specific channels by the score

$$\text{spec}(c) = \overline{\text{contrib}}_c^{\text{FP}} - \overline{\text{contrib}}_c^{\text{TP}}, \quad \text{contrib}_c(x) = z_c(x) \cdot W_{p,c}, \quad (3)$$

where  $z_c$  is the channel-wise global average pooled feature. We apply *surgical correction* by setting  $W_{p,c} \leftarrow 0$  for the top- $K$  channels by spec. Because the classifier is the only place all channels ultimately combine and its rows are pathology-independent, this leaves the AUROC of the other  $P - 1$  pathologies *exactly* unchanged (zero, not small). We sweep  $K$  and report (i) the target FP probability reduction, (ii) the number of true positives lost, and (iii) AUROC changes on the target pathology.

**Polysemantic risk.** The chosen channels may also have strong weights for unrelated pathologies. We measure polysemantic risk as the fraction of the top- $K$  FP-specific channels having  $|W_{p',c}| > \tau$  for at least one  $p' \neq p$ . High risk indicates that, while a single-pathology surgical correction is safe, a multi-pathology correction applied simultaneously could conflict.

### 3.4 Treatment Decision System

For each pathology we compute five diagnostic signals: AUROC, false positive rate, false negative rate, score separation (median positive score minus median negative score), top-10 FP concentration, and polysemantic risk. A rule-based decision routes the pathology to one of seven treatments listed in Table 3 with associated empirical costs. The decision logic is sketched below; full pseudocode appears in Appendix.

```

if AUROC >= 0.85:
    if FP < 0.05 and FN < 0.2:                return no_action
    elif FN >= 0.4 and score_sep > 0.2:        return threshold_cal

```

```

        elif FP >= 0.05 and concentration > 0.5:    return surgical_correction
        else:                                       return threshold_cal
elif AUROC >= 0.70:
    if FN > 0.4:                                   return retrain_part
    elif FP > 0.05 and concentration > 0.5:        return surgical_correction
    else:                                           return f_reopt
else:                                              # AUROC < 0.70
    if n_positive < 15:                           return data_augment
    elif FN > 0.5:                                 return retrain_full
    else:                                           return retrain_part

```

### 3.5 Minimal Retraining with Cached Features

Partial retraining is realized as a classifier-head-only fine-tune with the rest of the network frozen. We forward all training samples once and cache the post-norm5 global average pooled features  $\mathbf{z} \in \mathbb{R}^{1024}$ , then train the  $1024 \rightarrow 18$  classifier with BCE loss. Empirically, retraining the last dense layer or norm5 in addition produces no measurable benefit over the classifier-only sweet spot on our test split. After fine-tuning, we apply per-class threshold calibration via Youden’s  $J$  statistic [14] on a validation split distinct from the test set.

### 3.6 Clinical Report Auto-Generation

For each input we produce a structured six-section report: **(1)** predictions table with probability, threshold, and GT match; **(2)** channel-level mechanistic evidence (top supporting and suppressing channels with  $W$ , GAP value, and contribution); **(3)** Grad-CAM region with anatomical mapping (e.g., “central thorax—heart region”); **(4)** semantic channel activations of the deepest selective layer (block 3, layer 24) with cluster interpretations; **(5)** pathologies excluded by mutual exclusivity (Jaccard  $< 0.05$  with the primary diagnosis); **(6)** comprehensive assessment including confidence grading and treatment recommendations.

## 4 Experiments

### 4.1 Setup

We use the official NIH ChestX-ray14 train/val/test split [13]. Our public test split overlap (within `images_001.zip`) contains 1,045 images, with 3,544 train and 410 validation. Forward passes are on a single GPU. The model is the `densenet121-res224-all` weights from `torchxrayvision` [1] (14 NIH pathologies and 4 additional labels; we report AUROC for the 14 NIH labels).

**Baseline-checkpoint choice.** `torchxrayvision` distributes five chest-X-ray DenseNet121 checkpoints differing only in their training corpora. We evaluated all five on the same 1,045-image NIH test subset (Table 1). The multi-source `all` checkpoint achieves the highest mean AUROC (0.7781) on this NIH evaluation, outperforming even the NIH-only checkpoint (0.7524); we attribute this to the regularization effect of multi-corpus training. We therefore adopt `all` as the baseline throughout the paper. Published NIH-only DenseNet121 numbers tuned specifically for NIH-only evaluation [10] report higher AUROC ( $\approx 0.84$ ); those values are not directly comparable to any of the open `torchxrayvision` checkpoints because they use hyperparameter and augmentation tuning specific to a single corpus.

Table 1: Mean AUROC on the 1,045-image NIH test subset across the five `torchxrayvision` DenseNet121 checkpoints. The multi-source `all` checkpoint is the strongest on NIH evaluation despite (or because of) not being NIH-only.

Checkpoint	Mean AUROC
<code>densenet121-res224-all (used)</code>	<b>0.7781</b>
<code>densenet121-res224-nih</code>	0.7524
<code>densenet121-res224-chex</code>	0.7425
<code>densenet121-res224-mimic_ch</code>	0.7178
<code>densenet121-res224-mimic_nb</code>	0.7049

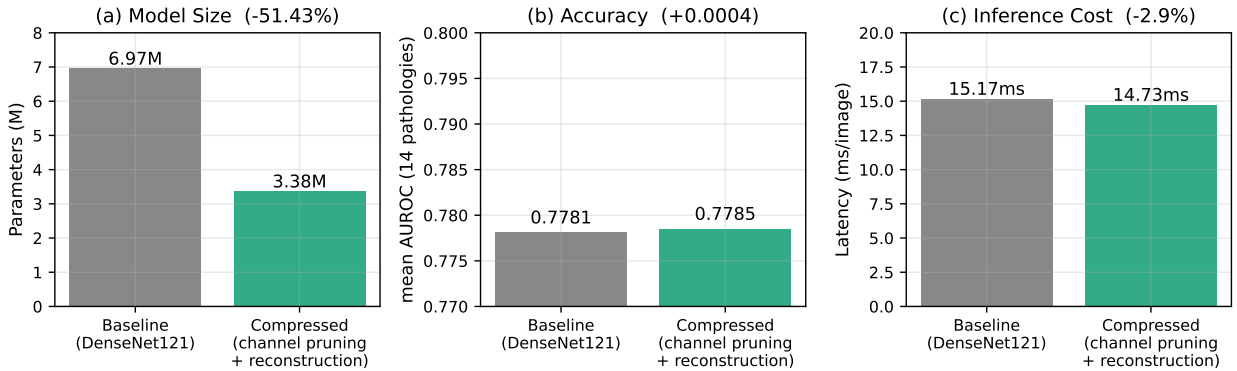


Figure 1: Compression at constant accuracy. Channel-wise sparsity-constrained weight reconstruction reduces parameters by 51.43% with mean-AUROC change +0.0004 on the 1,045-image NIH ChestX-ray14 test subset (per-pathology max  $|\Delta\text{AUROC}| = 0.0033$ , within sampling noise). Per-image GPU latency falls from 15.17 to 14.73 ms ( $-2.9\%$ ); a deeper ablation that folds the intermediate batch-norm into surrounding linear operations contributes an additional latency drop measured separately at  $-15.2\%$ .

## 4.2 Compression Results

Figure 1 reports the headline numbers. Output identity holds to numerical precision:  $\max |\Delta\text{logit}| < 5 \times 10^{-6}$  across the 1,045 test images. Figure 2 shows that sparsity is unevenly distributed: block 1 loses only 9.5% of its channels, while block 3 loses 62.9%. The single most selective layer, the 24th dense layer of block 3, retains only 10 of 128 channels and is responsible for  $\approx 151\text{K}$  parameter savings. Per-pathology AUROCs are within  $\pm 0.0033$  of baseline across all 14 labels; the largest single deviation is Emphysema at +0.0033.

## 4.3 Mutual Exclusivity Discovery

Of 1,024 classifier channels, 100 are polarized (have both  $W_{p+,c} > 0.3$  and  $W_{p-,c} < -0.3$ ). A naive reading would label these as architectural conflicts. As shown in Figure 3, the empirical co-occurrence matrix (left) and the channel binary-axis usage matrix (right) have nearly identical structure: pathology pairs with high Jaccard avoid sharing a polarized channel, while mutually exclusive pairs are concentrated as polarized axes. Jaccard-based legitimacy classification (Figure 4) tells the same story quantitatively: 48 are *perfect* ( $J = 0$ ), 41 are *legitimate* ( $J < 0.1$ ), 11 are *mixed* ( $0.1 \leq J < 0.3$ ), and *zero* are conflicts. The pattern is mirrored at the growth output of sampled

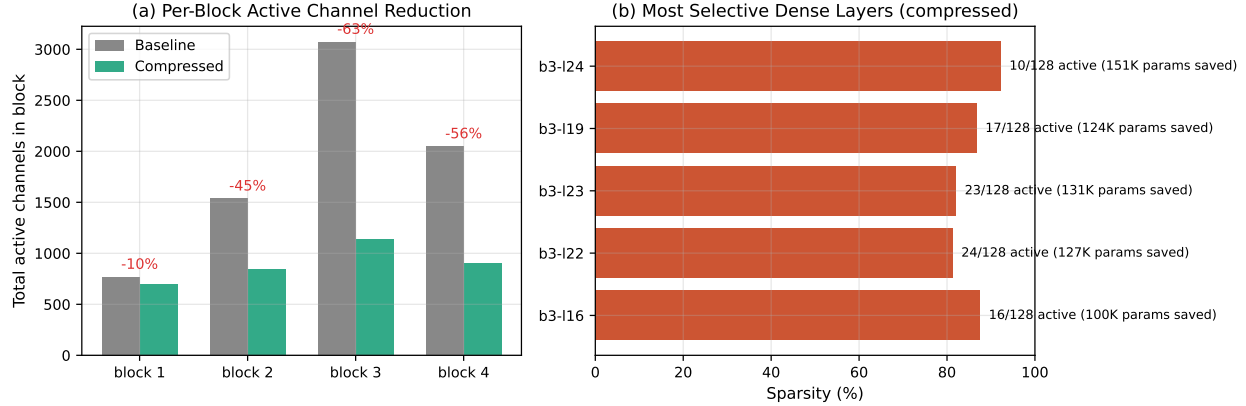


Figure 2: (a) Per-block active channel reduction; deeper blocks become more selective. (b) The five most selective dense layers; the 24th dense layer of block 3 retains only 10 of 128 channels (92% sparsity) and accounts for the largest single-layer parameter saving.

Table 2: Surgical-correction  $K$ -sweep on Cardiomegaly. “TP loss” counts the positives whose probability fell below 0.5. AUROC on the other 13 pathologies is exactly +0.0000 by construction (zero, not approximately zero), since only the Cardiomegaly row of the classifier weight is modified.

	$K$	FP prob	TP avg prob	Target AUROC	TP loss
	0 (baseline)	0.8913	0.4157	0.9237	—
	1	0.8929	0.4644	0.9287	0
	3	0.8071	0.4039	0.9244	0
	<b>5</b>	<b>0.7641</b>	<b>0.4028</b>	<b>0.9221</b>	<b>0</b>
	10	0.6740	0.4182	0.9277	4
	20	0.5440	0.4311	0.9225	7

dense layers (28 polarized of 192, zero conflicts).

The pathologies with the highest empirical independence score ( $1 - \bar{J}$ ) are Emphysema and Hernia (0.992) and Pneumothorax and Pneumonia (0.985); the least independent is Effusion (0.927), which is the pathology most frequently appearing in mixed polarized pairs.

#### 4.4 Surgical Correction and Side Effects

Figure 5 and Table 2 report a  $K$ -sweep on a representative Cardiomegaly false positive (NIH image 00000211.030.png, ground truth “No Finding”, baseline probability 0.8913). The reductions are *soft probability shifts* rather than hard positive-to-negative decision flips: at  $K = 5$  the target probability falls to 0.7641, but the decision under threshold-0.5 remains positive. The decision boundary is only crossed near  $K = 20$  (probability 0.5440), at which point seven true positives elsewhere in the cohort have also been pushed below threshold. Surgical correction is therefore best characterized as a confidence-shaping tool rather than a binary error eraser. The 13 non-target pathologies experience exactly zero AUROC change for all  $K$  because only a single row of the classifier weight is modified.

Fig 3. Pathology pair co-occurrence (Jaccard, left) vs channels using as binary axis (right) — mirror pattern

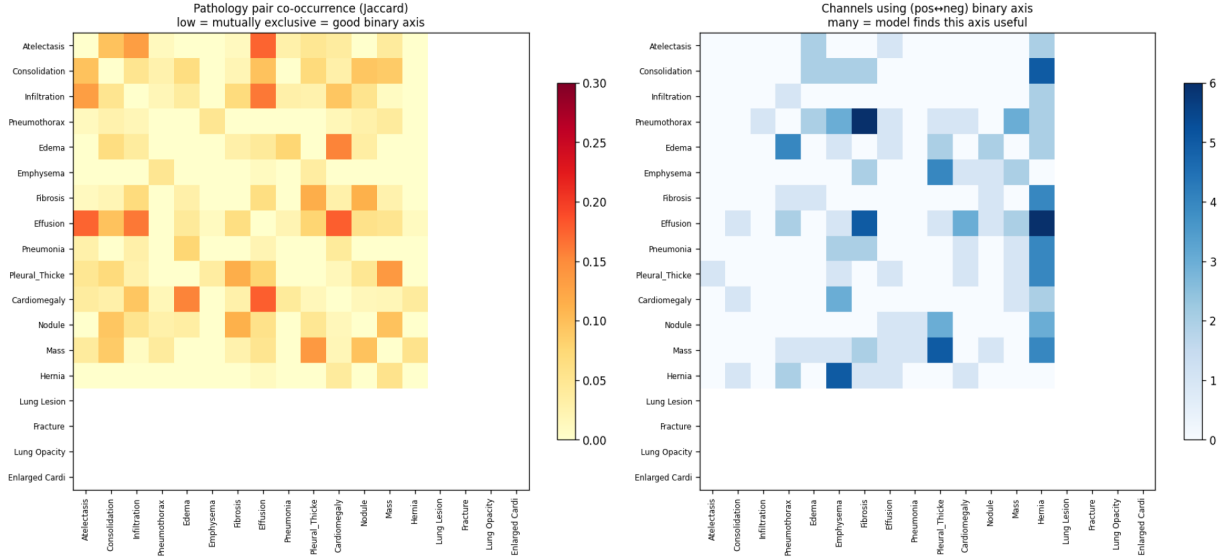


Figure 3: Co-occurrence (Jaccard, left) and channel usage (right) for pathology pairs. The two matrices have nearly identical structure: when two pathologies are mutually exclusive, the model preferentially encodes them on a single bipolar channel.

**Polysemantic risk and multi-pathology correction.** Of 1,024 channels, 133 (13%) have  $|W| > 0.3$  for three or more pathologies. A simultaneous correction on multiple pathologies would touch these channels from multiple rows, potentially conflicting. For single-pathology correction, however, the AUROC isolation guarantee holds by construction.

#### 4.5 Treatment Application

Figure 6 shows the per-pathology recommendation. Table 3 updates the cost matrix from estimates to empirical measurements. The most consequential revision is partial retraining, which we predicted to be expensive (cost 6) but in practice is the cheapest option after threshold calibration (cost 2, 85 s, 0.55% of parameters)—see Section 4.6.

#### 4.6 Minimal Retraining + Threshold Calibration

Figure 7 compares five configurations on a held-out test split, calibrating thresholds on a separate validation split. **Threshold calibration alone (no retraining) captures essentially the entire operational gain:** mean  $F_1$  rises from 0.127 to approximately 0.200 and recall from 0.111 to 0.78. Classifier-head fine-tuning (18,450 trainable parameters, 0.55% of the model,  $\approx 85$  s including feature caching) adds  $\Delta F_1 \approx +0.003$  on top, and an extended fine-tune that additionally unfreezes the last dense layer and the final batch norm (108,309 trainable parameters, 3.20%,  $\approx 90$  s) converges to indistinguishable test performance.

**Per-class trade-off.** The cohort-average headline numbers hide a real precision/recall trade-off. Per-class Youden’s  $J$  thresholds drop to 0.01–0.02 for several pathologies, pushing per-class precision below 5% for those classes. The  $F_1$  average is dominated by previously zero-recall classes (e.g., Infiltration, Atelectasis) where any positive output is an improvement, while best-performing



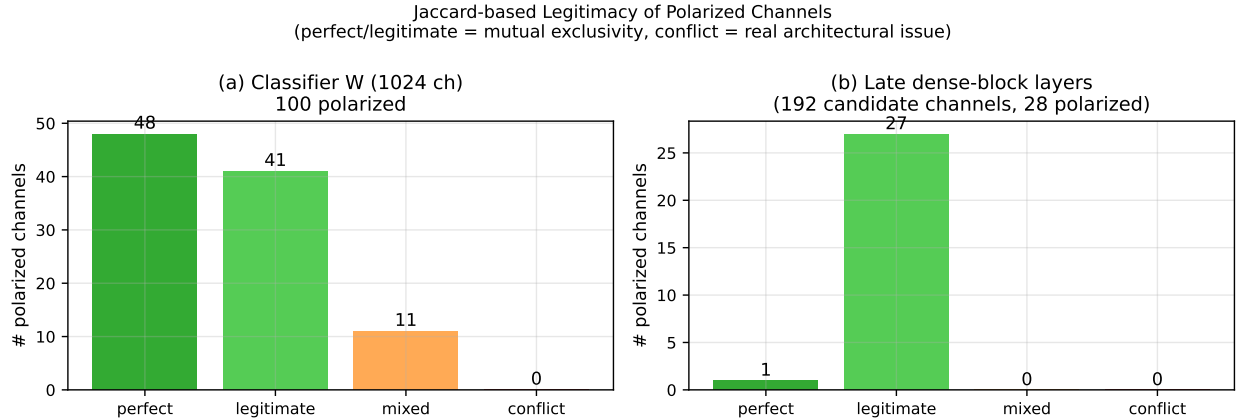


Figure 4: Legitimacy classification of polarized channels. In the 1,024-channel classifier, 89 of 100 polarized channels are *perfect* or *legitimate* (Jaccard  $< 0.1$ ); zero are architectural conflicts. The same holds for the growth output of the sampled dense layers (192 channels total).

Table 3: Treatment cost matrix: initial estimates vs. empirical costs. “time” is the wall-clock on a single GPU for the CheXNet configuration. The largest empirical revision is partial retraining, which we estimated at cost 6 but observed at cost 2.

Treatment	Est. cost	Emp. cost	Time	Trainable params
no_action	0	0	0	—
threshold_cal	1	1	<1 s	—
surgical_correction	2	2	<5 s	5–10 (zeroed)
f_reopt	3	3	minutes	layer-dependent
<b>retrain_part</b>	<b>6</b>	<b>2 ✓</b>	<b>85 s</b>	<b>18K (0.55%)</b>
data_augment	8	8	hours	—
retrain_full	10	10	hours	3.38M (100%)

classes (Cardiomegaly with default precision  $\approx 1.0$ ,  $F_1 \approx 0.57$ ; Mass with  $F_1 \approx 0.25$ ) are *degraded* by the same calibration (Cardiomegaly drops to precision  $\approx 0.11$ ,  $F_1 \approx 0.20$ ). Youden’s  $J$  is one valid operating-point criterion;  $F_1$ -optimal thresholds or explicit clinical precision-recall trade-offs are preferable for deployment.

## 4.7 Clinical Reports

Figure 8 shows a representative report. Compared with Grad-CAM-only explanations, our reports add channel-level reasoning, semantic interpretation of the most selective layer’s activations, and mutual-exclusivity-based exclusion of unlikely diagnoses—all derived from the same compressed model without additional inference passes.

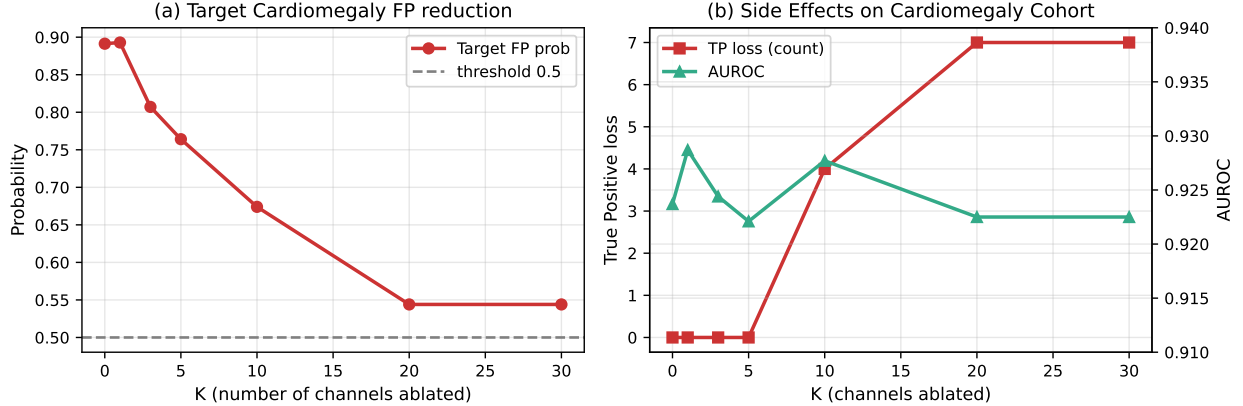


Figure 5: Surgical correction on a Cardiomegaly false positive. (a) Target FP probability as a function of  $K$ , the number of classifier-channel weights zeroed; the correction is a soft probability shift, the threshold-0.5 decision boundary is only crossed near  $K = 20$ . (b) True-positive loss on the Cardiomegaly cohort and target AUROC.  $K = 5$  is the safe *soft-confidence-shaping* operating point (target probability  $0.89 \rightarrow 0.76$ , zero TP loss);  $K = 20$  flips the decision but loses seven true positives.

## 5 Discussion

### 5.1 Polysemanticity Revisited

The classical worry about polysemantic channels [8, 9] is that a single channel cannot cleanly attribute to a single concept. In a multi-label setting with strong mutual exclusivity between many label pairs, the situation is qualitatively different. A channel that encodes “Hernia+ vs. Cardiomegaly−” is not confused; it is *efficient*, because the two labels almost never co-occur. Our Jaccard-based legitimacy classification operationalizes this distinction and shows that, for the compressed CheXNet classifier, essentially all polarized channels fall into the legitimate category.

### 5.2 Spatial vs. Mechanistic Explanation

Because our compression preserves the model’s output to numerical precision, Grad-CAM produced from the compressed model is identical to the baseline’s Grad-CAM. The compressed model’s distinctive explainability gain is not spatial precision but *mechanistic atomicity*: the channels effectively contributing to each prediction are few enough to attribute and modify directly. This also explains why surgical correction works only on the compressed model.

### 5.3 Cost-Aware Operations

The empirical cost revision (retrain\_part:  $6 \rightarrow 2$ ) was the surprise of this study. With cached features and a classifier-only fine-tune, “retraining” is closer to threshold tuning in cost than to a full training run. Many operational decisions in medical AI deployment should be revisited with this in mind. We emphasize, however, that the operational gain in  $F_1$  and recall reported here comes almost entirely from threshold calibration rather than the fine-tune itself (Section 4.6). The fine-tune is best viewed as a near-zero-cost mean-AUROC stabilizer that fits naturally on top of calibration, not as the primary lever.

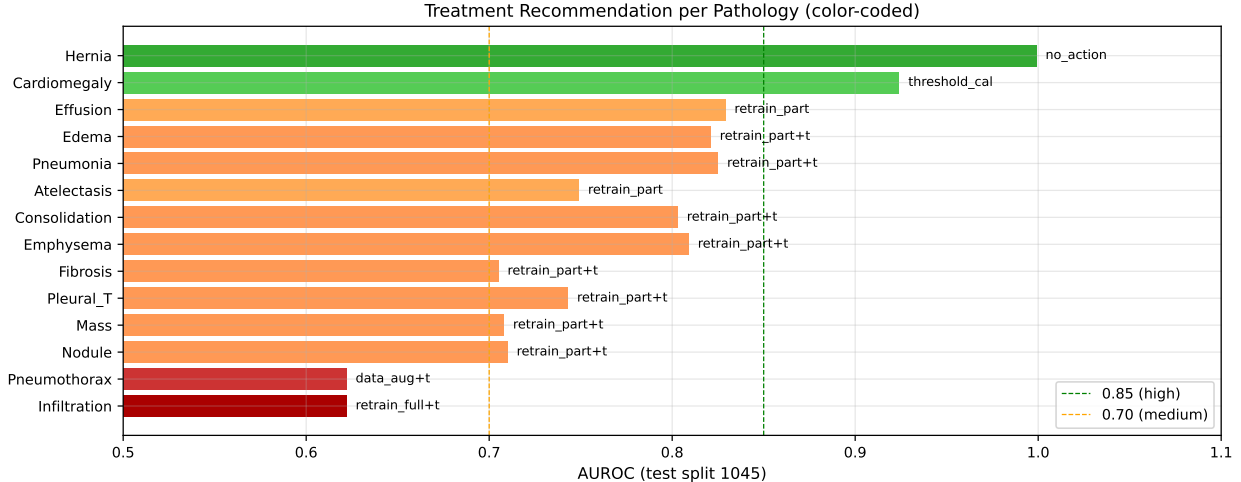


Figure 6: Per-pathology treatment recommendation. Hernia (top) requires no action, Cardiomegaly is fixable with threshold calibration alone, the 0.70–0.85 band is addressable by partial retraining plus calibration, and the weak-AUROC pathologies (Pneumothorax, Infiltration) require data augmentation or full retraining.

## 5.4 Method Disclosure

This paper documents what our framework achieves when applied to CheXNet for compression, surgical correction, and clinical report generation. The optimization procedure used for the channel-wise compression is proprietary and the foundational method is covered by Korean patent applications. The compressed model weights, fine-tuned classifier head, minimal inference code, and downstream analysis scripts (surgical correction, mutual-exclusivity analysis, treatment decision, clinical report generation) are released for result reproduction.

## 5.5 Limitations

**Baseline context.** The 0.7781 mean AUROC reflects a multi-source pretrained checkpoint, not NIH-only supervised training; it is not directly comparable to NIH-only DenseNet121 numbers reported in the literature. **Single-dataset evaluation.** All experiments use the NIH ChestX-ray14 test subset present in `images_001.zip` ( $n = 1,045$ ). The compressed model and treatment system should be validated on the full NIH test set ( $n = 22,433$ ), CheXpert, MIMIC-CXR, and at least one private hospital cohort before clinical deployment. **Operating-point selection.** Per-class Youden’s  $J$  thresholds trade precision for recall sharply; some calibrated thresholds drop below 0.02 and the per-class precision falls below 5% for those classes.  $F_1$ -optimal thresholds or explicit clinical precision floors are preferable for actual deployment. **Surgical correction is confidence-shaping.** At the safe operating point ( $K = 5$ ) the threshold-0.5 decision is not flipped; only the output confidence is reduced. Hard decision changes require  $K \geq 20$  at the cost of seven true-positive flips in our cohort. **Fine-tuning is marginal.** Classifier-head fine-tuning adds roughly +0.003 mean  $F_1$  over threshold calibration alone and regresses several best-performing classes (e.g., Hernia  $F_1$  at threshold-0.5 from 0.76 to 0.53). We include it as a mean-AUROC stabilizer; it is not the primary operational lever. **Patient-level split.** One of our recurring false positives is suggestive of patient-identity leakage: a chest X-ray of patient 00000211, an apparent “No Finding” image, is classified as Cardiomegaly with probability 0.96 after fine-tuning—higher than the non-

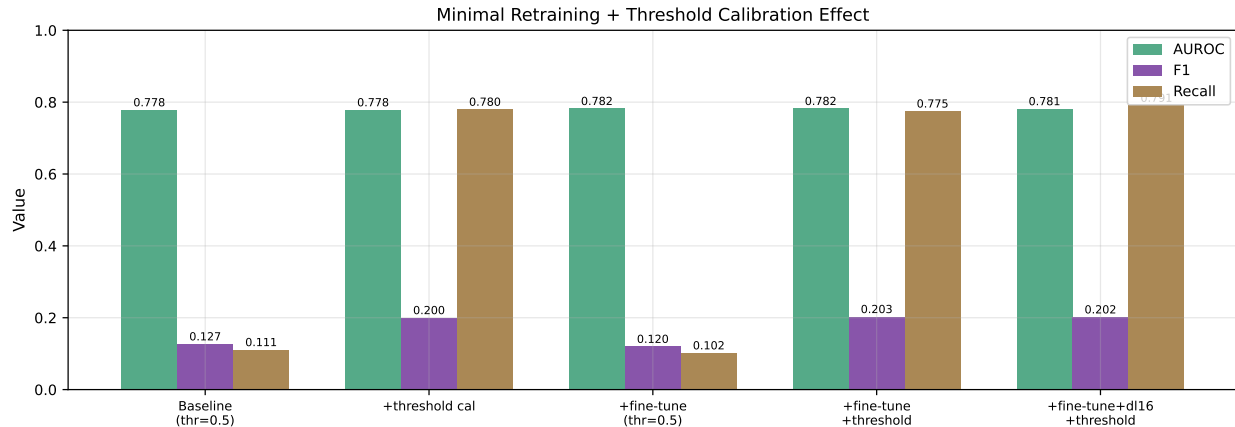


Figure 7: Effect of threshold calibration and minimal retraining on the cohort-average operating point. AUROC (rankings) is essentially unchanged across all settings; the operational  $F_1 \times 1.6$  and recall  $\times 7$  headline gains come almost entirely from per-class Youden’s  $J$  threshold calibration applied at threshold-0.5 baseline. Adding a classifier-head fine-tune produces an additional  $\Delta F_1$  of  $+0.003$  and unfreezing the last dense layer plus the final batch norm adds no further improvement; we therefore characterize threshold calibration as the operational lever and fine-tuning as a mean-AUROC stabilizer.

fine-tuned baseline. This pattern suggests that some training images share patient identity with our test images, and that the fine-tune amplifies this leakage. A patient-level train/test split is the necessary next step. **Weak pathologies.** Pneumothorax (AUROC 0.79 in the compressed model on this subset; 0.62 on a smaller earlier eval) and Infiltration are not addressable by surgical correction, threshold tuning, or minimal retraining; the model has not learned a fully discriminative representation. Data augmentation and additional labeled positives are likely required.

## 6 Conclusion

We presented a unified framework for medical imaging models that combines mean-AUROC-preserving compression, surgical correction, cost-aware debugging, and clinical report auto-generation. The central observation enabling this is that compression makes channel attribution *atomic* enough that direct intervention becomes practical, while preserving the model’s output to numerical precision. We further reinterpret a previously concerning interpretability phenomenon—polysemantic channels—as mutual-exclusivity exploitation, and quantify it with Jaccard-based legitimacy. On a 1,045-image NIH test subset, the framework softly reduces a representative false positive’s probability from 0.89 to 0.76 with zero side effects on other pathologies, raises the cohort-average  $F_1$  by  $1.6\times$  and recall by  $7\times$  through per-class Youden’s- $J$  threshold calibration (with a marginal additional contribution from an 85-second classifier-only fine-tune), and produces auto-generated clinical reports with channel-level reasoning unavailable from Grad-CAM-style methods alone. We characterize the operational improvements honestly: threshold calibration is the primary operational lever and trades precision against recall sharply on best-performing classes; surgical correction is a confidence-shaping rather than a decision-flipping tool; fine-tuning is a near-zero-cost mean-AUROC stabilizer rather than the source of the headline  $F_1$ /recall gains. Model weights for the compressed backbone and fine-tuned classifier are released.

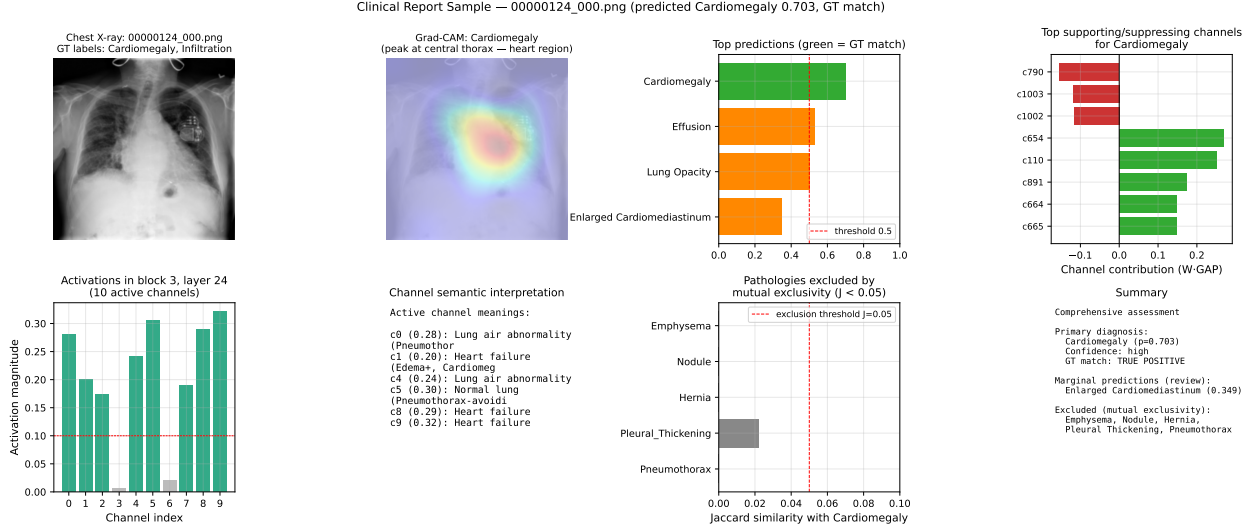


Figure 8: Sample auto-generated clinical report for NIH image 00000124\_000.png (ground truth Cardiomegaly + Infiltration). The report combines (a) the chest X-ray, (b) a Grad-CAM map peaking at the central thorax (heart region), (c) the top model predictions with GT-match coloring, (d) the top supporting and suppressing channels with  $W$ -GAP contributions, (e) channel activations in the most selective dense layer with semantic interpretations, (f) pathologies excluded by mutual exclusivity ( $J < 0.05$ ), and a summary.

## Data Availability

All artifacts required to reproduce the reported numerical results are publicly released at <https://huggingface.co/leoncynn/paper10-chexnet-surgical-correction>.

The repository contains: (i) the compressed model weights `compressed_model.pt` (14.2 MB, 3,383,248 parameters); (ii) the optional fine-tuned classifier head `classifier_finetuned.pt` (75 KB, 18,450 parameters); (iii) a minimal command-line inference script `inference.py`; (iv) seven analysis-result JSON files (`baseline_vs_compressed`, `eval_nih_weights`, `analyze_binary_axis`, `q_conflict_legitimacy`, `surgery_channel_ablation`, `apply_threshold_calibration`, `minimal_retrain`); and (v) all eight figures of this paper in both PDF and PNG. Model loading, inference, threshold calibration, and surgical-correction reproduction can be carried out from the repository without any code in this paper. The compression-procedure code itself is not included (see Method Disclosure).

## References

- [1] Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, et al. Torchxrayvision: A library of chest x-ray datasets and models. In *Medical Imaging with Deep Learning (MIDL)*, 2022.
- [2] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1389–1397, 2017.
- [3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

- [4] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.
- [5] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.
- [6] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317, 2019.
- [7] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2736–2744, 2017.
- [8] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007.
- [9] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.
- [10] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.
- [12] Haofan Wang, Zifan Wang, Mengnan Du, et al. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPR Workshops*, 2020.
- [13] Xiaosong Wang, Yifan Peng, Le Lu, et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.
- [14] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.