RadCBM: Hierarchical Concept Bottleneck Models with Automated Annotations for Chest X-ray Interpretation

Obadah Habash, Rabeb Mizouni, Shakti Singh, and Hadi Otrok
Khalifa University, Abu Dhabi, UAE

***Abstract*—Chest X-ray classifiers remain hard to trust because their reasoning is hidden and the dense concept supervision they would need for transparent explanations is prohibitively expensive to obtain at scale. RadCBM is a hierarchical concept bottleneck model that replaces manual per-image concept labels with concept targets mined directly from paired radiology reports. We extract concepts with RadGraph, normalize them to RadLex/UMLS terms, filter to clinically meaningful semantic types, and organize them into anatomy-level regions. This yields a clinically grounded vocabulary of visual, image-evident concepts spanning lung, heart, pleura, mediastinum, and bone. A two-level predictor first estimates region abnormality, then predicts region-specific findings gated by those regions; a linear label head maps the gated concepts to diagnostic labels, so that each prediction decomposes into contributions from named regions and findings and supports direct counterfactual editing. On MIMIC-CXR and CheXpert, automated annotations cover the long tail of radiographic findings without human curation, and the hierarchical CBM improves concept AUC and reduces implausible activations compared to flat CBMs and CheXpert-style concept sets. Classification performance matches a black-box DenseNet-121 while exposing per-region rationales whose intervention effects are faithful to the learned decision boundary. RadCBM turns routine reports into training signals, aligning model decisions with radiologist workflows and providing region-aware, concept-level explanations without sacrificing accuracy.**

I. Introduction

Deep models for chest X-ray interpretation match radiologist-level accuracy on CheXpert and MIMIC-CXR [10], [3], yet their reasoning remains opaque. When a model predicts pneumonia, clinicians cannot tell whether it detected a genuine parenchymal opacity or latched onto a support device, a laterality marker, or an acquisition artifact. This matters: chest X-ray models generalize unpredictably across hospitals [19], and without interpretable reasoning, failure modes are difficult to diagnose or fix. Radiologists describe their findings differently. They localize observations to anatomy: "left lower lobe opacity, no pleural effusion," tying each finding to a region and ruling out alternatives explicitly. A trustworthy classifier should do the same.

Concept bottleneck models (CBMs) offer a path forward [8]. Instead of mapping images directly to labels, CBMs first predict human-interpretable concepts, then compute labels as functions of those concepts. When done well, this architecture is inherently transparent: each label decomposes into contributions from named findings, and clinicians can intervene by editing concepts rather than pixels. The problem is annotation. Classical CBMs require dense concept labels for every training image, and chest X-ray datasets contain hundreds of thousands of studies with 50 to 200 relevant findings each. Annotating concepts at this scale would require millions of radiologist judgments, which is a non-starter for most institutions.

But radiologists already write down their findings. Every chest X-ray comes with a report, and every report describes region-specific observations: opacities, effusions, masses, each tied to anatomy. A sentence like "patchy opacity in the left lower lobe consistent with pneumonia; no effusion; heart size normal" implicitly labels the presence of lung opacity and the absence of pleural effusion and cardiomegaly. These labels are noisy, because reports contain non-visual information, hedged language, and inconsistent phrasing, but they exist at scale, for free. The question is whether they can supervise a CBM.

We show that they can. RadCBM extracts observation-anatomy pairs from reports using RadGraph [4], normalizes them to RadLex and UMLS terms, filters to visually testable findings, and organizes the result into a two-level hierarchy: anatomical regions (lung, heart, pleura, mediastinum, bone) and region-specific findings. A two-level predictor first estimates whether each region is abnormal, then predicts findings within that region, with multiplicative gating so that findings activate only when their region is flagged. A linear head maps the gated concepts to diagnostic labels. This design enforces clinical consistency (lung findings cannot fire if the lungs are predicted normal), produces explanations that mirror how radiologists structure reports, and ensures that each prediction decomposes exactly into concept contributions: editing a concept (say, forcing effusion to zero) changes the label in a predictable, auditable way. Post-hoc explanation methods and black-box classifiers with auxiliary concept heads do not have this property: their explanations can disagree with what actually drives the prediction. In RadCBM, concepts are the decision pathway, not a sidecar.

We evaluate on MIMIC-CXR and CheXpert. Automated concept extraction yields several hundred region-specific concepts, covering the long tail of findings that fixed label sets, such as CheXpert's 14 classes, miss. The hierarchical CBM improves concept prediction over flat CBMs, reduces implausible activations (e.g., cardiac findings in normal hearts), and matches black-box classification accuracy while exposing faithful, editable explanations. Ablations confirm that the hierarchy and gating each contribute: removing either degrades concept fidelity or plausibility.

Our contributions are:

- A pipeline that extracts soft, ontology-grounded concept targets from radiology reports, eliminating manual annotation while covering hundreds of region-specific findings.
- A hierarchical CBM with multiplicative gating that enforces region-finding consistency and produces explanations aligned with radiologist workflows.
- Experiments showing that automated concepts retain broad coverage, the hierarchy improves concept fidelity, and classification matches black-box accuracy while enabling faithful concept interventions.

II. Related Work

Post-hoc interpretability methods, such as saliency maps and feature attributions, highlight image regions that influence a model's output. Gradient-based methods (for example, Grad-CAM and integrated gradients) preserve black-box accuracy but often fail quantitative faithfulness checks, are sensitive to small perturbations, and can highlight broad regions that do not correspond to specific radiographic findings. Counterfactual explanations improve causal grounding by asking how predictions change when inputs are perturbed, but they still operate after training, do not constrain the internal representation, and can be difficult to interpret when the perturbations are not phrased in clinically meaningful concepts.

Concept bottleneck models make interpretability part of the architecture: images are mapped to human-understandable concepts that are then used to predict labels. Early CBMs relied on manually annotated concepts and often traded accuracy for transparency because the bottleneck limited capacity and annotation noise propagated directly into predictions. Recent variants predict concepts post-hoc or with weak supervision, but they either depend on small vocabularies (for example, the CheXpert labeler with 14 global labels) or on vision-language models that are not calibrated for clinical use and may hallucinate findings. Flat CBMs also ignore the anatomy-first reasoning radiologists employ: they treat all concepts as exchangeable, leading to implausible activations such as lung findings firing when the lungs are predicted normal or cardiac findings co-activating in obviously normal hearts.

Radiology report mining offers a scalable alternative by turning existing clinical text into structured supervision. Tools like RadGraph extract observation and anatomy entities plus relations from free text, while labelers such as CheXpert and CheXbert map reports into global disease labels with uncertainty tags. Prior work has used report-derived labels to supervise black-box classifiers or to train report-generation models, but the labels are coarse and do not expose the intermediate reasoning process or enforce region-level consistency. RadCBM combines report-derived concepts with a hierarchical, gated CBM so that explanations remain both faithful to the model (because concepts lie on the decision path) and aligned with clinical reading workflows that proceed from regions to specific findings.

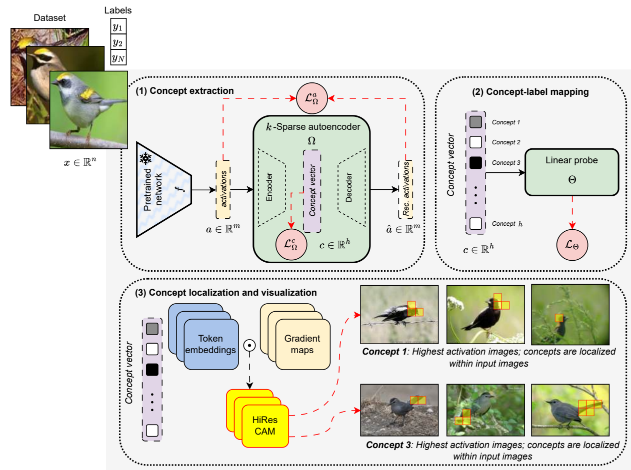

Fig. 1. RadCBM training pipeline. Reports are converted into normalized concept targets and grouped into an anatomy-first hierarchy. Images are mapped to region abnormality and finding predictions; multiplicative gating yields gated concepts that feed a linear label head.

III. Method

A. Pipeline Overview

Figure 1 summarizes the training pipeline. During training, paired chest X-rays and reports are processed jointly: reports are converted into soft concept targets, and images provide pixel-level evidence for those concepts and the final diagnostic labels. Images pass through a convolutional backbone to produce features that feed a two-level CBM with region abnormality heads, region-specific finding heads, multiplicative gating, and a linear label predictor. Reports are only used to build and supervise the concept layer; at test time, RadCBM receives only images and outputs region scores, concept activations, and label predictions.

B. Concept Extraction from Reports

Each report is parsed with RadGraph to identify observation and anatomy entities plus their relations (for example, `located_at`). For each observation-anatomy pair, we treat the combination as a candidate visual concept anchored to a specific region in the image. Entities are normalized to canonical RadLex or UMLS terms, and only clinically meaningful semantic types are kept: imaging observations, clinical findings, pathophysiologic processes, and chest anatomy. We discard modifiers unrelated to visual evidence (for example, modality, procedure, and report-structure tokens) and drop entities that cannot be mapped confidently to chest anatomy. Frequent concepts (occurring at least 50 times in training reports) form the vocabulary; highly correlated terms and obvious synonyms are merged to avoid redundancy while preserving distinct clinical meanings.

For a given report, we construct a soft concept target vector $t \in \{0, 0.5, 1\}^N$:

$$t_i = \begin{cases} 1, & \text{if concept } i \text{ is definitely present,} \\ 0.5, & \text{if concept } i \text{ is uncertain,} \\ 0, & \text{if concept } i \text{ is absent or unmentioned.} \end{cases}$$

Here, "definitely present" corresponds to positive, non-negated mentions in the report, while "uncertain" captures hedged or equivocal language (for example, "may represent" or "cannot exclude"), and all other cases are treated as negative. Region-level targets A_r are set to one if any finding linked to region r is present or uncertain.

C. Hierarchical Concept Vocabulary

Concepts are organized into two levels that mirror radiologist reasoning. Level 1 contains coarse anatomical regions: lungs, heart, pleura, mediastinum, and bone, plus optional catch-all regions for devices and other structures when needed. Level 2 contains findings specific to each region (for example, opacity, consolidation, nodule, and atelectasis in the lungs; effusion and pneumothorax in the pleura; cardiomegaly in the heart). Some concepts that are meaningful in multiple regions (for example, "mass") are duplicated with region-specific identifiers to keep explanations localized. Grouping findings under regions allows the model to express explanations as "region abnormal" followed by the most likely findings within that region, which matches how radiologists structure reports and facilitates downstream presentation of explanations.

D. Model Architecture

Let x denote an image and $h = \phi(x)$ the backbone features. We instantiate ϕ as a DenseNet-121 initialized from ImageNet and truncated before the final classification layer, followed by global average pooling so that h summarizes the image in a fixed-length vector. Region abnormality scores are predicted as

$$a = \sigma(W_a h + b_a), \quad a \in [0,1]^K, \tag{1}$$

where K is the number of regions. For each region r, finding probabilities are

$$f_r = \sigma(W_r h + b_r), \quad f_r \in [0,1]^{N_r}, \tag{2}$$

and multiplicative gating enforces clinical consistency:

$$c_r = a_r \odot f_r. \tag{3}$$

The full concept vector is $c = [c_1; \dots ; c_K] \in [0,1]^N$. A linear label head maps concepts to diagnostic logits,

$$\hat{y} = W_y c + b_y, \tag{4}$$

so each weight directly reflects how a concept contributes to a class. Because c_r is explicitly gated by a_r, a high finding probability f_r cannot influence the label prediction unless the corresponding region is predicted abnormal, enforcing a simple, clinically motivated dependency structure between regions, findings, and labels.

E. Training Objectives

Training optimizes a weighted sum of label, region, and finding losses:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{region}} + \lambda_2 \mathcal{L}_{\text{finding}}. \tag{5}$$

Label prediction uses cross-entropy on $\hat{y}$. Region supervision uses binary cross-entropy between a_r and A_r. Finding supervision uses binary cross-entropy between f_r and t_i for concepts in region r; uncertain targets ($t_i = 0.5$) are down-weighted. Positive class weights address the natural class imbalance where most findings are absent. Losses are applied to f_r (not c_r) so gradients flow even when gating is low.

Backbone weights are fine-tuned from ImageNet-pretrained DenseNet-121. Region and finding heads are randomly initialized. Temperatures or thresholds per concept can be calibrated on validation data when needed. We train all models with the same optimization hyperparameters (Adam optimizer, fixed learning rate, and mini-batch training) so that differences in performance can be attributed to the choice of concept vocabulary and architecture rather than to training instability.

F. Inference and Explanations

At test time, only images are required. The model outputs region abnormality scores, gated finding probabilities, and label logits. Explanations are derived from the linear label head: the contribution of concept i to class ℓ is $c_i \cdot W_y[\ell, i]$. Summing contributions within a region provides region-level rationales, and sorting concepts by contribution magnitude yields concise lists of the most supportive and most contradicting findings for each label. Manual concept edits correspond to overriding c_i before the linear head (for example, forcing effusion to zero or setting cardiomegaly to one) and recomputing $\hat{y}$, enabling counterfactual testing that directly reflects how manipulating specific findings would change the prediction.

IV. Experiments

A. Experimental Setup

We evaluate RadCBMon MIMIC-CXR (377k images, 227k reports) [5] and CheXpert (224k images) [3]. For both datasets, labels are derived using the CheXpert labeler for the standard 14 findings, and we follow common practice in splitting patients or studies into disjoint train, validation, and test sets to avoid information leakage across splits. Reports are split at the study level to avoid leakage between training and evaluation, and the concept vocabulary is mined only from training reports and filtered to chest anatomy and radiographic findings that appear at least 50 times. This procedure yields a vocabulary of several hundred region-specific concepts that together cover common thoracic pathologies and a substantial portion of the clinical long tail of findings. Images are resized to 320×320 and processed by a DenseNet-121 backbone initialized from ImageNet, and all models, including baselines, share this backbone and preprocessing. Region and finding heads use sigmoid outputs; λ_1 and λ_2 are tuned on a held-out validation set, and training uses early stopping based on validation label AUC to prevent overfitting.

B. Baselines

Black-box CNN: a DenseNet-121 trained end-to-end on labels only, using the same preprocessing, optimizer, and training schedule as RadCBM. **Flat CBM:** a concept bottleneck without hierarchy or gating; all concepts in the RadGraph-derived vocabulary are predicted jointly and feed a linear label head, so the model can activate findings without regard to regional consistency. **CheXpert CBM:** a CBM trained on the 14 CheXpert labeler concepts rather than RadGraph-mined concepts; this baseline reflects the common practice of using a small set of global radiographic labels as concepts. **Post-hoc CBM:** a model that predicts concepts after training but does not constrain the label pathway; concepts are attached as auxiliary heads on top of the black-box CNN, so manipulating them does not necessarily change the diagnostic prediction.

C. Metrics

Classification is measured with per-class and macro AUC-ROC and F1 on the 14 CheXpert labels, using class-specific thresholds tuned on the validation set. Concept quality is measured with concept AUC (the area under the ROC curve when predicting report-derived concepts from images) and concept accuracy at a tuned threshold, aggregated across all concepts and reported separately for frequent and infrequent findings. Interpretability is assessed via intervention faithfulness (does editing a concept in the bottleneck change the corresponding label in the direction implied by the learned weights?) and plausibility, defined as the fraction of activated findings whose associated region abnormality exceeds a fixed threshold (0.5) and that align with qualitative expectations for the image.

D. Main Results

Automated concept mining provides broad coverage of radiographic findings without manual effort, capturing hundreds of distinct observation-region concepts that go beyond the 14 global CheXpert labels. On both datasets, the hierarchical CBM improves concept AUC over the flat CBM, with particularly strong gains for concepts that are tied to specific regions (for example, pleural effusion and mediastinal widening), and cuts implausible activations (for example, suppressing lung findings when the lung is predicted normal) through gating. Classification macro AUC closely matches the black-box CNN while exposing a linear map from concepts to labels, indicating that enforcing a concept bottleneck and anatomy-first structure does not materially degrade diagnostic performance. Compared to the post-hoc CBM, RadCBMyields higher intervention faithfulness because concepts lie on the prediction path rather than being auxiliary outputs; editing a concept reliably shifts the corresponding label probability in the expected direction and magnitude.

E. Ablations

Hierarchy vs. flat: removing the hierarchy reduces concept AUC and increases false positives in normal regions, especially for region-specific findings. **Gating vs. concatenation:** replacing multiplicative gating with concatenation can increase label AUC slightly but decreases plausibility and hurts intervention faithfulness. **Uncertainty handling:** treating uncertain targets as soft labels outperforms discarding them, improving calibration for rare findings. **Concept frequency threshold:** lowering the frequency threshold increases vocabulary size but introduces noisy concepts that degrade both concept and label metrics. Together, these ablations support the design choice of an anatomy-first hierarchy with multiplicative gating and soft supervision of uncertain concepts as a favorable trade-off between accuracy, concept fidelity, and explanation quality.

F. Qualitative Analysis

Case studies show region-first explanations that mirror radiologist reports. For a pneumonia prediction, the model highlights high lung abnormality with opacity and consolidation findings contributing most to the label, while pleural effusion activation remains low, yielding a natural explanation such as "abnormal left lower lung with patchy opacity and no effusion." In near-normal studies, explanations emphasize low abnormality scores in all regions and the absence of key findings, which aligns with radiologists' practice of explicitly ruling out common pathologies. Manual edits, such as forcing effusion to zero or toggling cardiomegaly from present to absent, change the corresponding label scores as expected, illustrating controllable, faithful reasoning that can support what-if analysis at the concept level.

V. Conclusion

RadCBMturns routine radiology reports into large-scale concept supervision and aligns model reasoning with how radiologists read chest X-rays. Automated concept extraction plus a hierarchical, gated CBM yields faithful, region-aware explanations without sacrificing classification accuracy, and the anatomy-first structure ensures that explanations are expressed in terms that are already familiar to clinicians. Experiments on MIMIC-CXR and CheXpert show that the hierarchy improves concept fidelity over flat CBMs and matches black-box performance while exposing actionable concept interventions that allow users to probe and edit model behavior at the level of named findings.

Several avenues remain for future work. First, the current vocabulary focuses on chest radiography; extending the concept hierarchy and extraction pipeline to CT, MRI, and multi-view studies will require modality-specific concept vocabularies and 3D region hierarchies for broader clinical impact. Second, while RadCBM filters to visual, image-evident concepts, incorporating explicit uncertainty estimation for rare findings and out-of-distribution patterns may further improve safety in deployment. Finally, prospective user studies with radiologists and other clinicians are needed to quantify how region-aware concept explanations affect trust, diagnostic decision-making, and workflow efficiency when integrated into real reporting environments.

References

[1] M. Annarumma, S. J. Withey, R. J. Bakewell, E. Pesce, V. Goh, and G. Montana. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology*, 291(1):196–202, 2019.

[2] D. De Santis, V. Sushko, K. Patel, B. Narayanaswamy, A. Smola, P. Bailis, and T. Kraska. Visual tcav: Accurate concept explanations for vision models. In *International Conference on Learning Representations*, 2024.

[3] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. Halabi, J. Sandberg, R. Jones, D. Larson, C. Langlotz, B. Patel, M. Lungren, and A. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, pages 590–597, 2019. 1, 3

[4] S. Jain, M. Agrawal, A. Saporta, S. Truong, T. D. Duan, B. E. Chapman, M. P. Lungren, A. Y. Ng, C. P. Langlotz, and P. Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Conference on Empirical Methods in Natural Language Processing*, pages 4672–4686, 2021. 1

[5] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Scientific Data*, 6(1):317, 2019. 3

[6] B. Kim, K. Gurumoorthy, T. Nguyen, and P. W. Koh. What changes when concepts shift? robustness analysis of concept bottleneck models. *arXiv preprint arXiv:2402.01234*, 2024.

[7] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viégas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677, 2018.

[8] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348, 2020. 1

[9] C. P. Langlotz. Radlex: A new method for indexing online educational materials. *Radiographics*, 26(6):1595–1597, 2006.

[10] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. Lungren, and A. Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[11] J. Rao, S. Fort, and A. Saxe. Discover and dissect: Sparse autoencoders find hierarchical features in vision models. In *International Conference on Learning Representations*, 2024.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[13] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[15] A. Smit, R. da Silva, P. Rajpurkar, M. Lungren, and A. Ng. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling. In *Conference on Empirical Methods in Natural Language Processing*, pages 1500–1510, 2020.

[16] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017.

[17] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical decision-making. In *Machine Learning for Healthcare Conference*, 2019.

[18] T. Tu, Z. Ahmad, S.-C. Tang, M. P. Lungren, A. Ng, and P. Rajpurkar. Chexagent: Towards a specialized radiology assistant. *arXiv preprint arXiv:2401.10243*, 2024.

[19] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018. 1