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*—Abstract to be written towards the end...**

I. Introduction

Chest radiography remains the most frequently performed imaging examination worldwide, with hundreds of millions of studies acquired annually [1]. Interpreting these images is high-stakes: a missed pneumothorax, an overlooked nodule, or a mischaracterized cardiac silhouette can alter the trajectory of patient care [2]. Because the sheer volume of studies strains radiology workflows, diagnostic errors (while individually rare) accumulate into a substantial burden when multiplied across populations [3], [4]. The promise of computational assistance is therefore not merely academic. Systems that can reliably flag abnormalities, prioritize urgent cases, or provide differential considerations address a genuine clinical need [5].

Deep learning has delivered remarkable progress toward this goal. Convolutional and transformer-based architectures now match or exceed physician-level performance on curated benchmarks for thoracic pathology detection [6], [7], [8], [9]. These results, however, have not translated proportionally into clinical deployment [10], [11]. Part of this gap reflects concerns about robustness and generalization: models that perform well on internal test sets can fail when applied to external hospitals, sometimes because they exploit institution-specific artifacts rather than disease-related signal [12]. Reasons also include regulatory, infrastructural, and cultural barriers [10].

A recurrent critique in high-stakes clinical machine learning is that black-box predictions lack inspectable reasoning [13]. A model may assert "cardiomegaly" with high confidence, but it cannot articulate *why* or point to the cardiothoracic ratio it implicitly computed. It cannot situate its judgment within the anatomical and pathophysiological logic that governs clinical interpretation. This is not mere aesthetic preference for explanation. Radiologists think in concepts such as consolidation, air bronchograms, Kerley lines, and costophrenic blunting, and a system that cannot speak this language offers predictions without a basis for trust or correction [14].

Concept-based models, often instantiated as Concept Bottleneck Models (CBMs), offer an architectural response to this limitation [15]. Instead of mapping pixels directly to diagnostic labels, they introduce an intermediate representation of human-interpretable attributes. The model first predicts whether specific concepts are present (anatomical structures, radiographic findings, device positions) and then uses those concepts to produce diagnostic outputs. Explanations are thus part of the forward pass rather than added post hoc through saliency methods [16]. When a CBM predicts pulmonary edema, we can inspect whether it detected cardiomegaly, vascular redistribution, or interstitial opacities and check that this reasoning aligns with clinical knowledge.

This interpretability, however, comes at a cost that has limited practical adoption: concept-based models require concepts. Specifically, they require a predefined vocabulary of clinically meaningful attributes and, more demandingly, supervisory signal indicating which concepts are present in which images. Manual annotation at this granularity is expensive, time-consuming, and difficult to scale [17]. A single chest radiograph might exhibit dozens of relevant findings across multiple anatomical regions, each requiring expert assessment. Curated ontologies such as SNOMED CT provide standardized vocabularies [18], but these resources define *what* concepts exist, not *where* they appear in any particular image. The gap between the conceptual richness that would make these models clinically useful and the annotation budgets that real projects can sustain has constrained concept-based approaches to modest scales or narrow concept sets [19].

Recent attempts to apply concept-based models to medical imaging point toward two directions, neither yet fully satisfactory for chest radiography. The first generates concept vocabularies from large language models: one approach prompts GPT-4 to enumerate radiographic findings, then projects CLIP embeddings onto these concepts [20]. While this eliminates manual annotation, LLM-generated concepts lack grounding in clinical ontologies, may include findings that are not visually testable from a frontal radiograph, and inherit the hallucination tendencies of their source models. A second, natural direction is to repurpose existing extraction tools for supervision. Tools such as RadGraph [21] parse reports into entity–relation graphs and have become standard for evaluating report generation quality via RadGraph F1 scores, but, to our knowledge, have not been used to *supervise* concept bottleneck models. Their structured outputs remain confined to evaluation metrics rather than serving as trainable concept targets, and have not yet been developed into scalable concept banks for chest radiography. Meanwhile, routine radiology reports already encode rich conceptual supervision: by the time a radiologist documents "right basilar pneumonia," they have localized disease, described its radiographic pattern, and linked observations to a diagnostic impression. This information is

recorded in natural language rather than structured labels, but it is expert-generated, temporally aligned with the image, and available at scale in virtually every institution with an electronic health record [22]. The challenge is to turn this free text into supervision suitable for training concept-based vision models.

Transforming free-text reports into structured concept representations is not straightforward. Radiology language is dense with abbreviations, implicit negations, and context-dependent qualifications [23]. A finding may be "present," "absent," "unchanged," or "cannot be excluded," distinctions that matter clinically and must be preserved in any derived supervision [24], [25]. Linking extracted mentions to standardized terminologies introduces additional complexity: the same concept may be expressed in myriad surface forms, and disambiguation requires domain-specific knowledge. Recent advances in clinical natural language processing and biomedical entity linking [26], [21], together with resources such as the Unified Medical Language System (UMLS) [27], make such extraction increasingly tractable. However, these tools have rarely been combined into pipelines that produce *trainable* concept banks with assertion status, anatomical context, and ontological grounding.

This work addresses the gap between the latent supervision encoded in radiology reports and the structured representations that concept-based vision models require. Prior efforts have tackled adjacent problems: extracting findings from clinical text [28], [7], linking medical entities to ontologies such as UMLS [27], and training interpretable classifiers on manually curated concept sets [15]. These efforts, however, stop short of turning large report corpora into trainable, ontology-grounded concept banks and pairing them with hierarchical CBMs for chest radiography. Unlike systems that use reports primarily to derive noisy image-level labels for black-box classifiers or that restrict CBMs to small, hand-designed concept sets, we convert routine report corpora into ontology-grounded concept banks and use them to supervise RadCBM at the scale of institutional radiology archives.

On MIMIC-CXR and CheXpert, RadCBM matches the classification performance of strong black-box baselines while improving concept AUC and reducing implausible activations compared to flat CBMs. Automated annotations cover the long tail of radiographic findings without human curation, and the hierarchical architecture exposes region-aware rationales whose counterfactual edits faithfully track the learned decision boundary.

The contributions of this work are threefold:

- We introduce RadCBM, the first hierarchical concept bottleneck architecture for chest radiography that organizes concepts by anatomical region, gates region-specific findings through region abnormality scores, and constrains label predictions to linear functions of gated concepts. This design enforces clinical consistency (lung findings cannot fire when lungs are predicted normal) and produces explanations aligned with radiologist workflows.
- We present a framework that repurposes RadGraph, previously used only for report generation evaluation, as a source of trainable concept supervision. By linking extracted mentions to SNOMED CT via the UMLS and preserving assertion status, we construct ontology-grounded concept banks at scale without manual per-image annotation, covering hundreds of region-specific findings beyond the 14-class vocabularies typical of prior work.
- We provide empirical analysis on MIMIC-CXR and CheXpert demonstrating that RadCBM matches black-box classification accuracy while improving concept AUC over flat CBMs, reducing implausible activations through gating, and enabling faithful concept interventions whose effects reliably track the learned decision boundary.

The remainder of this paper is organized as follows. Section II situates our work within related efforts in chest radiograph analysis, concept-based modeling, and clinical natural language processing. Section X describes the concept extraction pipeline, from report preprocessing through entity linking to concept bank construction, and details the model architectures and training procedures for both concept prediction and downstream classification. Section X presents experimental results on large-scale chest radiograph datasets. Section X discusses limitations, clinical implications, and directions for future work.

II. Related Work

A. Deep Learning for Chest Radiography

Large-scale datasets have driven rapid progress in automated chest radiograph interpretation. ChestX-ray14 provided over 100,000 images with NLP-derived labels [29]; CheXpert [7] and MIMIC-CXR [22] expanded scale while improving label quality and providing associated reports. Architectures from DenseNet-based CheXNet and CheXNeXt [6], [30] to Vision Transformers [8], [9] now match radiologist performance on common pathologies. Clinical adoption nevertheless lags, partly because these models offer predictions without reasoning. Post-hoc explanations, including saliency maps [31] and Grad-CAM [16], show *where* models attend but not *what* they detect, failing to bridge the gap between neural activations and the conceptual vocabulary radiologists use [13].

B. Concept Bottleneck Models

Concept Bottleneck Models (CBMs) address interpretability by routing predictions through human-interpretable intermediate representations [15]. The model first predicts concept presence, then reasons from concepts to outputs, making the decision process transparent by construction. Extensions include post-hoc retrofitting of pretrained networks [32], concept embeddings that relax strict bottlenecks [33], and interactive variants enabling test-time correction [34]. Applications span dermatology [35], ophthalmology [36], and radiology. The persistent limitation is concept acquisition: training requires annotations for every concept, and manual labeling at the granularity needed for clinical utility is prohibitively expensive [19]. Ontologies define concept vocabularies but not their image-level presence.

Recent work has sought to reduce dependence on manual concept labels and to better characterize the faithfulness and robustness of concept-based explanations. Label-free CBMs and language-guided bottlenecks align CLIP-style vision-language representations with concept predictors, discovering

concepts and names without per-concept supervision [19], [37]. Visual TCAV and related approaches refine concept scoring and selection [38], while GlanceNets [39] and concept-shift analyses [40] highlight structural and robustness limitations, showing that concept pipelines can still exploit shortcuts even when their explanations appear plausible. Our approach is complementary: rather than discovering concepts from generic image-text corpora, we construct an ontology-grounded concept bank directly from radiology reports and use it as the bottleneck for chest X-ray interpretation. Critically, while RadGraph and similar tools have become standard for *evaluating* report generation systems via entity-level F1 scores [21], [21], they have not previously been used to *supervise* concept bottleneck models. Our work bridges this gap, converting RadGraph's structured extraction into trainable concept targets with assertion status and anatomical localization.

C. Clinical NLP for Radiology Reports

Radiology reports encode concept information in natural language, motivating automated extraction. Rule-based systems like NegBio [28] and the CheXpert labeler [7] match patterns to identify findings and their assertion status. CheXbert improved on these using BERT fine-tuned on expert annotations [25], and RadGraph extended extraction to full entity-relation graphs [21]. Assertion detection, distinguishing present, absent, and uncertain findings, remains critical, addressed by systems from NegEx [24] through modern neural classifiers [41]. These tools extract increasingly structured information from reports, though integration into pipelines producing trainable concept banks remains underdeveloped.

D. Biomedical Entity Linking

Grounding extracted mentions in standardized terminologies normalizes linguistic variation and enables semantic reasoning. UMLS integrates over 200 vocabularies, including SNOMED CT, into a unified metathesaurus [27]. Neural linking methods, particularly SapBERT's self-alignment pretraining on UMLS synonyms [26], achieve strong performance mapping surface forms to canonical concepts. This machinery enables extracted findings to be represented in ontology-grounded form suitable for concept-based modeling.

E. Vision-Language Models in Medical Imaging

Contrastive pretraining on image-text pairs offers an alternative path to leveraging reports. CLIP's success [42] prompted medical adaptations: ConVIRT [43], MedCLIP [44], and BiomedCLIP [45] align radiograph and report representations, enabling zero-shot classification through textual prompting. These approaches handle unpaired data and transfer flexibly across tasks. However, learned representations remain entangled rather than decomposed into discrete concepts, trading interpretable structure for representational flexibility [37].

The components for concept-based chest radiograph modeling, including clinical NLP, entity linking, concept architectures, and vision-language alignment, exist but remain fragmented. This work integrates them into a pipeline that produces structured concept banks from report archives, enabling concept-based modeling at institutional scale.

References

[1] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, and E. C. Rosenow, "Interpretation of plain chest roentgenogram," *Chest*, vol. 141, no. 2, pp. 545–558, 2012. 1

[2] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction," *Radiographics*, vol. 35, no. 6, pp. 1668–1676, 2015. 1

[3] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?" *Insights into Imaging*, vol. 8, no. 1, pp. 171–182, 2017. 1

[4] J. J. Donald and S. A. Barnard, "Common patterns in 558 diagnostic radiology errors," *Journal of Medical Imaging and Radiation Oncology*, vol. 56, no. 2, pp. 173–178, 2012. 1

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. 1

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017. 1, 2

[7] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597. 1, 2, 3

[8] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using vision transformers on chest x-rays," *Scientific Reports*, vol. 14, no. 1, 2024. 1, 2

[9] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023. 1, 2

[10] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, 2019. 1

[11] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins, and M. Maruthappu, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, p. m689, 2020. 1

[12] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Confounding variables can degrade generalization performance of radiological deep learning models," *arXiv preprint arXiv:1807.00431*, 2018. 1

[13] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. 1, 2

[14] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. von Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: challenges and opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190043, 2020. 1

[15] 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*, 2020, pp. 5338–5348. 1, 2

[16] 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*, 2017, pp. 618–626. 1, 2

[17] M. J. Willemink, W. A. Koszek, C. Tan, T. C. Defined, R. L. Defined, M. P. Defined, and B. N. Defined, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020. 1

[18] K. Donnelly, "Snomed-ct: The advanced terminology and coding system for ehealth," *Studies in Health Technology and Informatics*, vol. 121, pp. 279–290, 2006. 1

[19] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, "Label-free concept bottleneck models," in *International Conference on Learning Representations*, 2023. 1, 2, 3

[20] A. Yan, Y. Wang, Y. Zhong, Z. He, P. Karypis, Z. Wang, C. Dong, A. Gentili, C.-N. Hsu, J. Shang, and J. McAuley, "Robust and interpretable medical image classifiers via concept bottleneck models," *arXiv preprint arXiv:2310.03182*, 2023. 1

[21] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng *et al.*, "Radgraph: Extracting clinical entities and relations from radiology reports," in *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2021. 1, 2, 3

[22] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019. 2

[23] J. C. Denny, "Extracting structured information from free text: challenges and approaches," *AMIA Annual Symposium Proceedings*, vol. 2009, p. 161, 2009. 2

[24] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *Journal of Biomedical Informatics*, vol. 34, no. 5, pp. 301–310, 2001. 2, 3

[25] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 1500–1519. 2, 3

[26] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021, pp. 4228–4238. 2, 3

[27] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. suppl_1, pp. D267–D270, 2004. 2, 3

[28] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, "Negbio: a high-performance tool for negation and uncertainty detection in radiology reports," *AMIA Summits on Translational Science Proceedings*, vol. 2018, p. 188, 2018. 2, 3

[29] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017. 2

[30] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists," *PLoS Medicine*, vol. 15, no. 11, p. e1002686, 2018. 2

[31] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2014. 2

[32] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," in *International Conference on Learning Representations*, 2023. 2

[33] M. E. Zarlenga, P. Barbiero, G. Ciravegna *et al.*, "Concept embedding models: Beyond the accuracy-explainability trade-off," in *Advances in Neural Information Processing Systems*, 2022. 2

[34] K. Chauhan, R. Tiwari, J. Freyberg, P. Shenoy, and K. Dvijotham, "Interactive concept bottleneck models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 5948–5955. 2

[35] A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed, "On explainability of deep neural networks for medical image analysis," *arXiv preprint arXiv:2004.08780*, 2020. 2

[36] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visber *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018. 2

[37] Y. Yang, A. Panagopoulou, S. Sreekumar, I. Chalkidis, M. Yatskar, and C. Callison-Burch, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19 187–19 197, 2023. 3

[38] 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

[39] E. Marconato, A. Passerini, and S. Teso, "Glancenets: Interpretable, leak-proof concept-based models," in *Advances in Neural Information Processing Systems*, 2022. 3

[40] 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. 3

[41] A. Khandelwal and S. Sawant, "Negbert: A transfer learning approach for negation detection and scope resolution," *arXiv preprint arXiv:1911.04211*, 2020. 3

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763. 3

[43] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*, 2022, pp. 2–25. 3

[44] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3876–3887. 3

[45] S. Zhang, Y. Xu, N. Usuyama, J. Bagher, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong *et al.*, "Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023. 3