Can Vision-Language Models Enable More Efficient Concept-Based Learning with Less Supervision for Interpretable Lung Nodule Diagnosis?

Baoqiang Ma[1] B.MA-2@UMCUTRECHT.NL
Djennifer Madzia-Madzou[1] D.K.MADZIA-MADZOU-3@UMCUTRECHT.NL
Jin Ouyang[1] J.OUYANG@UMCUTRECHT.NL
Kenneth Gilhuijs[1] K.G.A.GILHUIJS@UMCUTRECHT.NL
[1] *Image Science Institute, University Medical Center Utrecht, the Netherlands.*

Editors: Under Review for MIDL 2026

Abstract

Interpretability is necessary for the safe deployment of AI systems in clinical practice, especially in tasks such as the diagnosis of lung nodules. Concept Bottleneck Model (CBMs) provide a promising framework for interpretable predictions by linking decisions to clinically meaningful concepts. However, standard CBMs rely on extensive and time-consuming concept annotations. Recent methods aimed to fill this gap by leveraging vision-language models (VLMs) for few-shot or even label-free concept learning. However, it still remains unclear whether the prior knowledge within VLMs is sufficient for fine-grained, nodule-level concept detection. In this work, we comprehensively investigate how much supervision is essential for reliable concept-based diagnosis, and whether VLMs can improve efficiency. We compare black-box models, standard CBMs, few-shot VLM-based CBMs, and label-free CBMs on CT-based lung nodule diagnosis. The results show that few-shot VLM-based CBMs achieve improved concept detection (Balanced accuracy (Bacc): 0.78 vs. 0.76, F1 score: 0.76 vs. 0.72) and diagnostic performance (Bacc: 0.72 vs. 0.52, 0.74 vs. 0.36) compared to standard CBMs, and can even outperform black-box models in F1 score (0.74 vs. 0.66). In contrast, label-free CBMs produce unreliable meaningless concept representations. These results suggest that VLMs can reduce supervision and improve interpretability and diagnostic performance, but are not yet sufficient for fully label-free concept-based learning.

Keywords: Concept learning, interpretability, lung nodule diagnosis, Vision-Language model.

1. Introduction

Reliable clinical application of AI radiology tools requires not only strong predictive performance but transparent decision-making. Concept-based approaches, including Concept Bottleneck Models (CBMs) (Koh et al., 2020), aim to extract clinically meaningful concepts from images and use them for prediction. These models have been explored in CT-based lung nodule analysis on the LIDC-IDRI dataset (Armato III et al., 2011), which includes annotations of interpretable concepts such as spiculation and texture. However, a primary limitation of CBMs is their reliance on expensive and time-consuming concept annotations from radiologists.

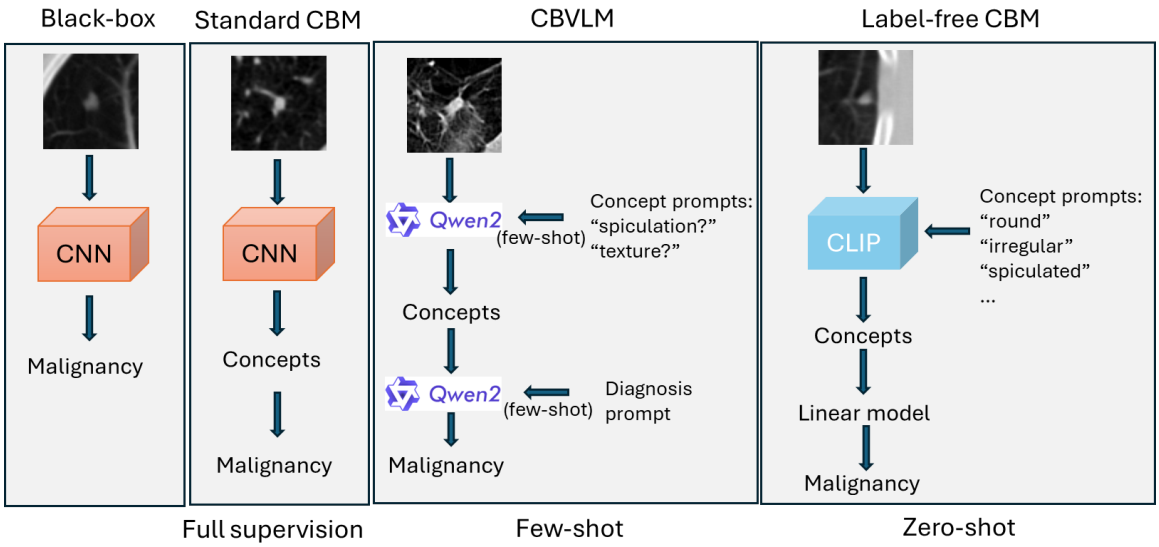

Figure 1: Overview of the comparison framework. We evaluate concept-based models under different supervision levels, including fully supervised CBMs, few-shot VLM-based CBMs (CBVLM), and zero-shot label-free CBMs.

This burden could potentially be addressed by recent progress in vision-language models (VLMs). VLMs such as CLIP (Radford et al., 2021) enable few-shot or zero-shot (label-free) concept learning by leveraging prior image–text understanding acquired from large-scale pretraining. Recent works have explored VLM-based concept models in few-shot setting (CBVLM) (Patrício et al., 2025), and have shown promising results across multiple medical imaging domains such as skin and chest X-ray imaging. In parallel, zero-shot approaches such as label-free CBMs (Oikarinen et al., 2023), typically built on CLIP, attempt to infer concepts without supervision and have shown strong performance in natural image domains. These approaches have also been extended to medical imaging tasks, such as skin and fundus image analysis (Chowdhury et al., 2024). However, these datasets involve well-defined and visually distinctive concepts, making them easier to detect than subtle, fine-grained nodule-related concepts in CT. It remains unclear whether such approaches generalize to more complex lung nodule characterization tasks.

This work aims to address this gap by investigating: (1) whether VLMs can reliably support concept learning for lung nodule analysis and (2) how much supervision is required to achieve effective and interpretable performance.

2. Materials and Methods

2.1. Data and Preprocessing

The public LIDC-IDRI dataset was used for lung nodule analysis. For each nodule, a 3D region of size $64 \times 64 \times 64$ mm^3 was cropped around the nodule center. The central axial slice was extracted and resized to 224×224 as model input. Malignancy labels and eight

Table 1: Comparison of different models for concept detection and malignancy prediction.

	Concept Detection		Malignancy Prediction	
Model	Bacc	F1	Bacc	F1
Black-box	–	–	**0.83**	0.66
CBM (Full)	0.76	0.72	0.52	0.36
CBVLM (Few-shot)	**0.78**	**0.76**	0.72	**0.74**
Label-free CBM	–	–	0.79	0.62

radiologist-annotated concepts (e.g., spiculation, texture) were binarized into $\{0, 1\}$. The dataset was split into training, validation, and test sets, containing 1197, 211, and 213 nodules, respectively.

2.2. Models

We compare four settings, as shown in Figure 1 : (1) a black-box baseline and (2) a standard CBM, which first predicts concepts and then uses them for diagnosis, both implemented with a 2D ResNet50 backbone. (3) A few-shot CBVLM, where concept predictions are obtained using a vision-language model (Qwen2) with in-context examples and text prompts (e.g., "spiculation", "texture"), and subsequently used for nodule diagnosis via an additional prompt-based inference step. (4) A zero-shot label-free CBM, where concept scores are derived from a CLIP model using predefined concept prompts and used for downstream linear malignancy prediction.

3. Results

Table 1 shows the performance of all methods. CBVLM achieved better concept detection performance compared to the fully supervised CBM(Balance accuracy (Bacc): 0.78, F1: 0.76). For malignancy prediction, the black-box model achieved the highest balanced accuracy (0.83), while CBVLM obtains the best F1 score (0.74), indicating the better balance between precision and recall. The label-free CBM showed competitive balanced accuracy (0.79) but lower F1 score, suggesting itsless stable predictions.

4. Discussion

Although Black-box model achieved strong predictive performance, but their lack of interpretability limits clinical use.Standard CBM achieved reasonable concept detection performance. However, its malignancy prediction is much worse than other models, likely due to misalignment between learned concepts and the target label.

CBVLM performs well in both concept detection and diagnosis. This is probably due to prior visual knowledge in VLMs and the use of few-shot guidance, which helps adapt the model to lung nodule analysis. In contrast, label-free CBMs fail to learn meaningful concepts in the experiments. This could be caused by no supervision and limited nodule-level concept knowledge in current VLMs.

Overall, future work should focus on combining few-shot learning into VLMs or training VLMs with detailed nodule-level concept annotations.

References

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Townim F Chowdhury, Vu Minh Hieu Phan, Kewen Liao, Minh-Son To, Yutong Xie, Anton van den Hengel, Johan W Verjans, and Zhibin Liao. Adacbm: An adaptive concept bottleneck model for explainable and accurate diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2024.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Cristiano Patrício, Isabel Rio-Torto, Jaime S Cardoso, Luís F Teixeira, and João C Neves. Cbvlm: Training-free explainable concept-based large vision language models for medical image classification. *Computers in Biology and Medicine*, 198:111145, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.