

Clinical Data-Driven Retrieval-Augmented Model for Lung Nodule Malignancy Prediction

Ruibo Hou¹, Shurong Chai¹, Rahul Kumar JAIN¹, Yinhao Li¹, Jiaqing Liu¹,
Shiyu Teng¹, Xiaoyu Shi¹, Lanfen Lin², and Yen-Wei Chen¹✉

¹ Graduate School of Information Science and Engineering, Ritsumeikan University,
Osaka, Japan

chen@is.ritsumei.ac.jp

² College of Computer Science and Technology, Zhejiang University, Hangzhou, China

Abstract. Deep learning techniques have been widely applied to lung nodule malignancy prediction tasks. Recently, the emergence of Vision-Language Models (VLMs) has enabled the use of textual information, further improving diagnostic accuracy. Nevertheless, two key limitations persist: (1) the insufficient utilization of clinical data to enhance computer-aided diagnosis, and (2) the limited ability of existing frameworks to leverage similar cases in the diagnostic process. To address these issues, we propose a clinical data-driven, retrieval-augmented VLM framework for lung nodule malignancy prediction. The proposed framework comprises a multimodal encoder, a retrieval-augmented module, and a text encoder. Lesion classification is achieved by evaluating the similarities between the combined visual and clinical data features and the text features of predefined categories, thereby establishing a robust mechanism for malignancy prediction. Moreover, the retrieval-augmented module further refines the prediction process by incorporating similar cases retrieved using clinical data as a query, thus facilitating more informed and accurate decisions. Overall, this framework comprehensively utilizes clinical data by integrating it into CT image features and enabling cross-interaction in the retrieval-augmented module to support diagnosis with similar cases. Experimental results on the publicly available LIDC-IDRI dataset demonstrate that the proposed framework achieves significant improvements in lung nodule malignancy prediction, with an approximate 3% increase in accuracy. Our code is released on Github: <https://github.com/chenn-clear/ClinicalRA>.

Keywords: Lung nodule classification · vision-language model · retrieval-augmented method.

1 Introduction

Lung cancer remains one of the most prevalent cancers worldwide, with the classification of lung nodules in chest CT images being a critical step for early detection and improved survival rates. In recent years, deep learning [1, 2, 9, 10] has been extensively applied to this task, significantly advancing the performance of automated diagnostic systems. Early approaches typically framed

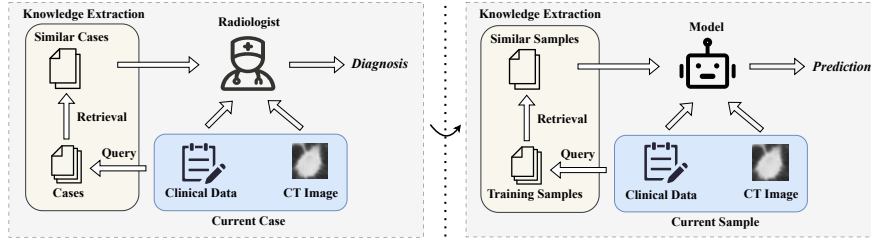


Fig. 1: Inspired by the diagnostic practices of radiologists, who rely on past similar cases for diagnosis, we propose a retrieval-augmented approach that enhances model predictions by leveraging similar samples.

lung nodule malignancy prediction as a binary classification problem [9, 16], distinguishing between benign and malignant cases. To better capture the gradual progression of disease severity, several studies have explored ordinal regression [4, 20] in medical imaging [11, 14], and reformulated the malignancy prediction task as a three-class classification problem by introducing an intermediate “unsure” category to account for diagnostic ambiguity.

Despite these advancements, significant challenges remain. A major issue is the underutilization of clinical data provided by radiologists, which often contains highly abstract and semantic-rich information capable of differentiating nodules with similar visual appearances but distinct pathological outcomes. To address this issue, CLIP-Lung [8] employed a Vision-Language Model (VLM) based on CLIP [13] to align textual representations of categories, visual features of CT images, and text information derived from clinical data. This approach demonstrated the potential of leveraging clinical data to capture abstract morphological features of lung nodules, thereby further improving diagnostic accuracy.

In clinical practice, radiologists often reference past similar cases to enhance diagnostic accuracy. Apart from above discussed methods, recent advances in retrieval-augmented approaches from Natural Language Processing (NLP) have begun to be adopted in VLM tasks. For example, methods such as RA-CLIP [15], RECO [6] and RA-CM3 [17] align visual and text features by retrieving similar samples based on image or text similarity. Nevertheless, this strategy remains underexplored in lung nodule classification. To address this gap, we propose a retrieval-augmented approach (Fig. 1) that integrates clinical data with CT images while concurrently retrieving similar cases to inform predictions. In contrast to image-driven retrieval methods, our approach mirrors physician reasoning by harnessing the semantic richness of clinical data to identify relevant samples. The core innovation of our framework lies in the comprehensive utilization of clinical data, enhancing image embeddings and guiding the retrieval process to provide evidence-based support for predictions. The contributions of this work are as:

1. We introduce a retrieval-augmented framework designed for lung nodule malignancy prediction. This framework mimics the diagnostic reasoning of radiologists by retrieving similar samples, enhancing predictive capabilities.
2. We propose a novel approach for retrieving similar samples using clinical data as a query. Our method comprehensively utilizes clinical data in two complementary ways: integrating it with visual features to enrich the image representation, and employing it as a query to identify the most semantically relevant samples.
3. Experimental results on the public LIDC-IDRI [3] dataset demonstrate significant improvements over previous methods. Notably, our approach excels in predicting the challenging intermediate “unsure” category, addressing a common weakness in earlier models.

2 Methodology

The proposed framework is depicted in Fig. 2. Inspired by CLIP-Lung [8], we adopt a VLM framework to align visual and clinical embeddings of lung nodule CT images with the textual representations of predefined categories. Unlike CLIP-Lung, which directly aligns clinical data, images, and categories, our approach employs a retrieval-augmented framework to enhance robustness and supplement physician-like knowledge extraction.

Our framework improves visual representations via two strategies: (1) clinical data is integrated into the multimodal encoder through a lightweight MLP, thereby enriching the visual features with semantic insights; and (2) clinical data is used as a query to retrieve relevant samples, offering rich contextual references that refine predictions. This dual use of clinical data mirrors physicians’ practice of combining current case details with historical examples for accurate diagnosis.

2.1 The Proposed Framework: An Overview

In our framework, each lung nodule CT image I is paired with corresponding clinical data D . Both images and corresponding clinical data are provided as input into the VLM framework. The VLM framework incorporates a multimodal encoder to process imaging and clinical data, a retrieval-augmented module, and a text encoder.

Multimodal Encoder. The detailed structure of the multimodal encoder is presented in Fig. 2(a). Following [8], we employ a ResNet-18 [5] to extract visual features from the CT image I , resulting in an embedding $h_v \in \mathbb{R}^{d_v}$. Meanwhile, values of attributes from clinical data D form a vector $h_d \in \mathbb{R}^{d_d}$. Here, v in h_v stands for “vision” and d in h_d stands for “data”. The visual and clinical embeddings are concatenated to form a unified representation:

$$h_v^d = \text{MLP}([h_v, h_d]) \quad h_v^d \in \mathbb{R}^{d_v}, \quad (1)$$

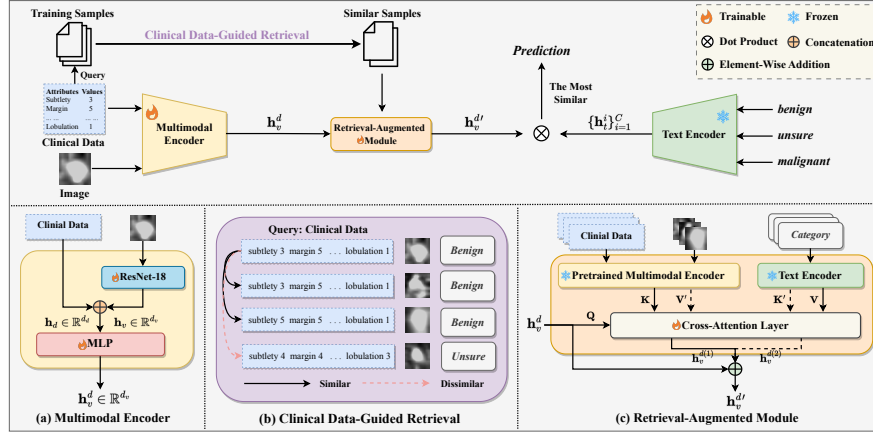


Fig. 2: The framework of the proposed method, showing an overview in the upper part and key components in the lower part.

where $[h_v, h_d]$ denotes the concatenation operation. The MLP layer transforms the combined embedding h_v^d , capturing both visual and clinical information while aligning its dimensionality (\mathbb{R}^{d_v}) with the predefined category text embeddings for similarity computation. This module is referred to as the multimodal encoder.

Retrieval-Augmented Module. This module comprises a frozen multimodal encoder, a frozen text encoder, and a trainable cross-attention layer, as shown in Fig. 2(c). Given an input sample, the retrieved top- K similar images $\{I_i\}_{i=1}^K$, along with their corresponding clinical data $\{D_i\}_{i=1}^K$ and category labels $\{y_i\}_{i=1}^K$, are fed into the module. The frozen pretrained multimodal encoder extracts visual-clinical embeddings $\{e_v^d\}_{i=1}^K$ from these samples, while the frozen pretrained text encoder obtains text embeddings $\{e_t\}_{i=1}^K$ (where t stands for “text”) for the category labels. The retrieved visual-clinical features interact with the visual-clinical embedding of the input sample (denoted as h_v^d) via key, query, and value projections, effectively integrating the retrieved information. The final output is a refined representation $h_v^{d'}$. Further details are provided in §2.2.

Text Encoder. The text embeddings $\{h_t^i\}_{i=1}^C$ corresponding to the predefined categories $\{c_i\}_{i=1}^C$, where C is the number of categories, are derived from a pretrained frozen text encoder (BioLinkBERT-large [18]). The final prediction is then made by calculating the cosine similarity between these text embeddings and the refined visual-clinical embedding $h_v^{d'}$.

2.2 Clinical Data-Driven Retrieval-Augmentation Training Scheme

The model training comprises two stages. In the **first stage**, we train a multimodal encoder that integrates the input image I and clinical data D , resulting in

an enriched visual-clinical representation h_v^d . This representation is then directly aligned with the text embeddings $\{h_t^i\}_{i=1}^C$ of the categories $\{c_i\}_{i=1}^C$, which are obtained from a frozen pretrained text encoder. The alignment is performed by computing the cosine similarities between the visual-clinical representation and the text embeddings. The prediction \hat{y} is calculated as follows:

$$p(\hat{y} = i) = \frac{\exp\left(\frac{\text{sim}(h_v^d, h_t^i)}{\tau}\right)}{\sum_{i'=1}^C \exp\left(\frac{\text{sim}(h_v^d, h_t^{i'})}{\tau}\right)}, \quad (2)$$

where $\text{sim}(h_t^i, h_v^d) = \frac{h_t^i \cdot h_v^d}{\|h_t^i\| \|h_v^d\|}$ denotes the cosine similarity, and τ adjusts the sharpness of the similarity distribution. The multimodal encoder is optimized using cross-entropy loss between predictions and ground truth. Once pretrained, it serves as a frozen backbone in the retrieval-augmented module for encoding retrieved samples and providing contextual references for the final prediction.

In the **second stage**, we first retrieve similar cases based on clinical data similarity, instead of the conventionally used image embedding similarity. Specifically, as presented in Fig. 2(b), we take the clinical data D as a query to retrieve the top- K most similar samples. The motivation lies in the fact that CT images can be visually similar yet differ significantly in their pathological attributes. In contrast, clinical data provides consistent and abstract information, making it more reliable for identifying relevant cases. This retrieval process is performed offline and accelerated using FAISS [7].

Next, inspired by RA-CLIP [15], we train a new multimodal encoder from scratch, while the pretrained encoders in the first stage are incorporated within a retrieval-augmented module. The retrieved samples $\{I_i, D_i, y_i\}_{i=1}^K$ are processed by the frozen pretrained encoders, whereas the primary inputs, image I and its associated clinical data D , are fed into the new multimodal encoder. The retrieval-augmented module then refines the main input representation by integrating knowledge from the retrieved samples, yielding an enhanced feature $h_v^{d'}$ that integrates both original and contextual information. As illustrated in Fig. 2(c), the visual-clinical embeddings $\{e_v^d\}_{i=1}^K$ and text embeddings $\{e_t\}_{i=1}^K$ from the retrieved samples are processed via a trainable cross-attention layer. The visual-clinical embedding of main inputs I and D obtained from the new multimodal encoder being trained, denoted as h_v^d , serves as the query in both cross-attention steps: first, h_v^d attends to $\{e_v^d\}_{i=1}^K$, with $\{e_t\}_{i=1}^K$ providing contextual information:

$$h_v^{d(1)} = \text{CrossAttention}\left(h_v^d, \{e_v^d\}_{i=1}^K, \{e_t\}_{i=1}^K\right). \quad (3)$$

Second, $h_v^{d(1)}$ attends to $\{e_t\}_{i=1}^K$, with $\{e_v^d\}_{i=1}^K$ contributing complementary visual context:

$$h_v^{d(2)} = \text{CrossAttention}\left(h_v^{d(1)}, \{e_t\}_{i=1}^K, \{e_v^d\}_{i=1}^K\right). \quad (4)$$

Finally, the enriched embeddings $h_v^{d(1)}$ and $h_v^{d(2)}$ are combined with h_v^d via element-wise addition to produce the refined representation:

$$h_v^{d'} = h_v^d + h_v^{d(1)} + h_v^{d(2)}. \quad (5)$$

The refined embedding $h_v^{d'}$ is then used to compute similarities with categories, following the procedure from Eq. (2), to generate the final prediction.

3 Experimental Settings

Dataset and Evaluation. All experiments are conducted on the publicly available LIDC-IDRI dataset [3], which contains low-dose CT images from 1,010 patients and serves as a benchmark for lung nodule classification tasks. All nodules with annotated malignancy scores ranging from 1 to 5 were extracted. Following previous works [8, 14], we categorized nodules based on their average malignancy scores as follows: nodules with scores below 2.5 are classified as benign; nodules with scores between 2.5 and 3.5 are considered unsure; nodules with scores above 3.5 are classified as malignant. Each nodule is accompanied by clinical data provided by radiologists, which includes eight distinct attributes, with each attribute having a corresponding value, that capture the key semantic characteristics of the nodules (e.g., sphericity, margin, lobulation, etc.). The nodules and their corresponding annotations, including clinical attributes, were extracted using the `pyl IDC` toolkit³. Following the preprocessing approach in [8], we cropped all nodules into a square-shaped volume, with a size corresponding to twice the equivalent diameter centered on the annotated location. The cropped nodules were then resized to a uniform volume of 32^3 voxels to ensure consistency across samples.

To ensure robustness and reliability, we conduct five-fold cross-validation experiments. The evaluation metrics include accuracy (%), which reflects overall classification performance, as well as recall and F1-score for each individual category (benign, unsure, and malignant) to provide a more comprehensive assessment of model performance across different nodule types.

Implementation Details. The model is trained with a batch size of 512, a learning rate of 0.01, and the SGD optimizer with a weight decay of 0.00005. Both training stages are conducted for 1,000 iterations each. Experiments were conducted efficiently on an NVIDIA RTX A6000 GPU, with approximately 4GB of memory usage and a total runtime of about 30 minutes for retrieval, training, and inference. Additionally, based on internal observations during model development, to balance retrieval quantity and quality, the number of retrieval samples K is set to 5 across all experiments.

4 Experimental Results and Analysis

Comparison with Other Methods. The comparative results with SOTA methods are presented in Table 1. Compared to ordinal classification methods (e.g., Poisson, NSB, UDM, and CORF) and VLM-based methods (e.g., CLIP,

³ <https://pyl IDC.github.io/>

Table 1: Classification results on the LIDC-IDRI dataset.

Method	Accuracy	Benign		Unsure		Malignant	
		Recall	F1	Recall	F1	Recall	F1
Linear Classifier	47.7±1.5	49.1	48.7	52.9	48.2	37.7	44.5
Poisson [4]	52.7±0.7	60.5	56.8	41.0	44.1	58.4	58.7
NSB [11]	53.4±0.7	80.7	63.0	16.0	24.2	67.3	63.8
UDM [14]	54.6±0.4	76.7	64.3	32.5	39.5	49.5	53.5
CORF [20]	56.8±0.4	71.3	63.3	38.5	44.3	61.3	62.3
CLIP [13]	56.6±0.3	59.5	59.2	53.9	52.2	55.2	60.0
CoCoOp [19]	56.8±0.6	59.0	59.2	55.1	52.8	55.2	60.0
CLIP-Lung [8]	60.9±0.4	67.5	64.4	53.4	54.1	60.9	66.3
Ours	63.8±1.6	66.4	68.5	60.9	58.8	64.1	64.2

Table 2: Ablation study on the contributions of clinical data and retrieval query to classification accuracy.

Encoder Input		Retrieval Query		Accuracy
Image	Clinical Data	Image Embedding	Clinical Data	
✓				52.6
✓	✓			57.6 (+5.0)
✓		✓		52.1 (-0.5)
✓	✓	✓		55.5 (+2.9)
✓			✓	62.4 (+9.8)
✓	✓		✓	63.8 (+11.2)

CoCoOp, and CLIP-Lung), our method achieves the highest accuracy. In the classification of all three categories, our approach performs remarkably well, underscoring the effectiveness of our retrieval-augmented framework in improving diagnostic robustness. Notably, it demonstrates a significant improvement in the “unsure” category, illustrating the advantage of leveraging retrieved samples to disambiguate ambiguous cases.

Ablation Study. Table 2 shows that adding clinical data to the main input improves accuracy from 52.6% to 57.6%, highlighting its semantic value. Using pretrained image embeddings as retrieval queries degrades performance due to irrelevant or inconsistent information misaligned with the pathology. In contrast, our clinical data-guided retrieval significantly improves accuracy, particularly for image-only inputs (by 9.8%). The best result (63.8%) is achieved when clinical data serves as both auxiliary input and retrieval query, highlighting its complementary role.

To better understand the observed differences, we compare top-1 retrieval accuracy using two types of queries. Image embeddings yield about 43% accuracy,

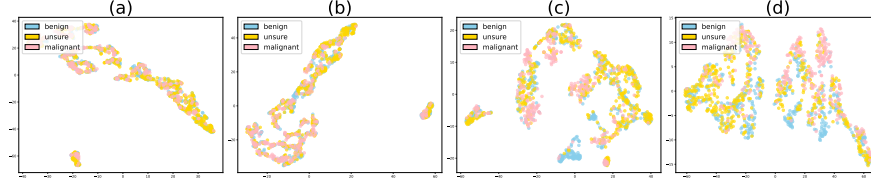


Fig. 3: The t-SNE results. (a) The base VLM; (b) VLM with clinical data-enriched multimodal encoder; (c) VLM with clinical data-guided retrieval augmentation; (d) Our final proposed method.













Query Type	Top-1	Top-2	Top-3	Top-4	Top-5	Final Prediction	Ground Truth
N/A	(a) No Retrieval (Base Model)						Unsure
	N/A	N/A	N/A	N/A	N/A	Benign	
	(b) Image-Based Retrieval						
	Benign	Benign	Benign	Benign	Unsure	Benign	Unsure
							
	(c) Clinical Data-Guided Retrieval						
	Unsure	Benign	Unsure	Benign	Unsure	Unsure	Unsure
							

Fig. 4: Qualitative comparison of an unsure case under three settings: (a) base model without retrieval, (b) image-based retrieval-augmented model, and (c) clinical data-guided retrieval-augmented model.

while clinical data achieve around 56%, highlighting a key limitation of image-based retrieval: its focus on surface-level features often misses subtle diagnostic cues. In contrast, clinical data provide high-level, diagnosis-relevant information, enabling the retrieval of cases more aligned with the underlying characteristics of lung nodules. Notably, directly using the label of the top retrieved sample yields 56% accuracy, which is both less rigorous and less effective than our proposed framework that integrates retrieval into a unified clinical-visual pipeline, achieving approximately 64%.

Feature Visualization and Case Study. Fig. 3 compares t-SNE[12] results across methods. While panels (b) and (c) focus on clinical data enrichment and retrieval, respectively, panel (d) integrates both, yielding distinct spatial distributions. In panel (d), malignant nodules (pink) cluster above the diagonal running from the top left to the bottom right, benign nodules (blue) lie below, and unsure nodules (yellow) fall in between. This separation highlights the framework’s ability to capture intrinsic relationships for accurate, interpretable classification.

Fig. 4 presents a case study of an unsure sample. Using clinical data as the retrieval query enables the model to access diagnostically relevant cases, leading to improved classification performance compared to the base model and image-based retrieval, particularly for this challenging ambiguous category.

5 Conclusion

We propose a retrieval-augmented framework for lung nodule malignancy prediction that integrates clinical data into the multimodal encoder and employs it as a retrieval query to identify more relevant samples. Experiments on the LIDC-IDRI dataset show significant improvements, especially in the challenging “unsure” category, validating the effectiveness of our approach. Future work will focus on validating the effectiveness and generalizability of the proposed framework on broader datasets.

Acknowledgments. This work was supported in part by the Grant in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant Nos. 20KK0234, 21H03470, and 20K21821, and in part by the National Key Research and Development Project (No. 2022YFC2504605).

Disclosure of Interests. The manuscript is approved for publication by all authors. All authors declare no conflicts of interest regarding the submission and publication.

References

1. Ai, Y., Liu, J., Li, Y., Wang, F., Du, X., Jain, R.K., Lin, L., Chen, Y.W.: SAMA: A self-and-mutual attention network for accurate recurrence prediction of non-small cell lung cancer using genetic and CT data. *IEEE Journal of Biomedical and Health Informatics* **29**(5), 3220–3233 (2025)
2. Al-Shabi, M., Shak, K., Tan, M.: Procan: Progressive growing channel attentive non-local network for lung nodule classification. *Pattern Recognition* **122**, 108309 (2022)
3. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Van Beeke, E.J., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Croft, B.Y.: 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)
4. Beckham, C., Pal, C.: Unimodal probability distributions for deep ordinal classification. In: *International Conference on Machine Learning*. pp. 411–419. PMLR (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
6. Iscen, A., Caron, M., Fathi, A., Schmid, C.: Retrieval-enhanced contrastive vision-text models. In: *The Twelfth International Conference on Learning Representations* (2024)

7. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* **7**(3), 535–547 (2019)
8. Lei, Y., Li, Z., Shen, Y., Zhang, J., Shan, H.: Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 403–412. Springer (2023)
9. Lei, Y., Tian, Y., Shan, H., Zhang, J., Wang, G., Kalra, M.K.: Shape and margin-aware lung nodule classification in low-dose ct images via soft activation mapping. *Medical Image Analysis* **60**, 101628 (2020)
10. Liao, Z., Xie, Y., Hu, S., Xia, Y.: Learning from ambiguous labels for lung nodule malignancy prediction. *IEEE Transactions on Medical Imaging* **41**(7), 1874–1884 (2022)
11. Liu, X., Zou, Y., Song, Y., Yang, C., You, J., K Vijaya Kumar, B.: Ordinal regression with neuron stick-breaking for medical diagnosis. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. pp. 0–0 (2018)
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* **9**(11) (2008)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
14. Wu, B., Sun, X., Hu, L., Wang, Y.: Learning with unsure data for medical image diagnosis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10590–10599 (2019)
15. Xie, C.W., Sun, S., Xiong, X., Zheng, Y., Zhao, D., Zhou, J.: Ra-clip: Retrieval augmented contrastive language-image pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19265–19274 (2023)
16. Xie, Y., Zhang, J., Xia, Y.: Semi-supervised adversarial model for benign-malignant lung nodule classification on chest ct. *Medical image analysis* **57**, 237–248 (2019)
17. Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., Yih, W.T.: Retrieval-augmented multimodal language modeling. In: *International Conference on Machine Learning*. pp. 39755–39769. PMLR (2023)
18. Yasunaga, M., Leskovec, J., Liang, P.: LinkBERT: Pretraining language models with document links. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 8003–8016. Association for Computational Linguistics, Dublin, Ireland (2022)
19. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022)
20. Zhu, H., Shan, H., Zhang, Y., Che, L., Xu, X., Zhang, J., Shi, J., Wang, F.Y.: Convolutional ordinal regression forest for image ordinal estimation. *IEEE transactions on neural networks and learning systems* **33**(8), 4084–4095 (2021)