Aligning Human Knowledge with Visual Concepts Towards Explainable Medical Image Classification

Yunhe Gao, Difei Gu, Mu Zhou, Dimitris Metaxas

Rutgers University

Abstract. Although explainability is essential in the clinical diagnosis, most deep learning models still function as black boxes without elucidating their decision-making process. In this study, we investigate the explainable model development that can mimic the decision-making process of human experts by fusing the domain knowledge of explicit diagnostic criteria. We introduce a simple yet effective framework, **Explicd**, towards **Exp**lainable **l**anguage-**i**nformed **c**riteria-based **d**iagnosis. Explicd initiates its process by querying domain knowledge from either large language models (LLMs) or human experts to establish diagnostic criteria across various concept axes (e.g., color, shape, texture, or specific patterns of diseases). By leveraging a pretrained vision-language model, Explicd injects these criteria into the embedding space as knowledge anchors, thereby facilitating the learning of corresponding visual concepts within medical images. The final diagnostic outcome is determined based on the similarity scores between the encoded visual concepts and the textual criteria embeddings. Through extensive evaluation of five medical image classification benchmarks, Explicd has demonstrated its inherent explainability and extends to improve classification performance compared to traditional black-box models. Code is available at `https://github.com/yhygao/Explicd`.

Keywords: Explainable Model · Vision Language Model · Visual Concept Learning

1 Introduction

The advent of deep learning [12,7] has profoundly transformed the field of medical image analysis [21,5,18,11] in lowering diagnostic costs [17,9,10] and improving diagnostic accuracy [26,2]. Despite these advancements, state-of-the-art deep neural networks often operate as black boxes. While they can achieve high performance in end-to-end image classification, they fail to provide the explainable rationale behind the decision-making process. This lack of transparency compromises the trust and validation by healthcare professionals, leading to potential errors and resistance of integrating AI-derived insights into clinical settings [6]. Unlike these black-box AI models, as illustrated in Fig. 1 (B), human experts make diagnoses by meticulously analyzing key image features from color, shape,

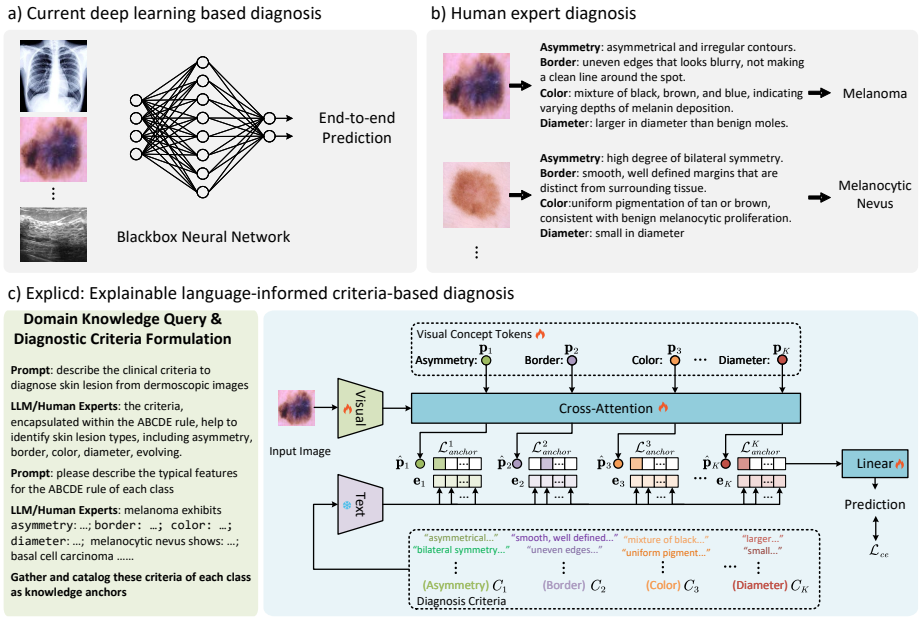

Fig. 1: (a) Current state-of-the-art deep learning models often function as black boxes, offering predictions without revealing insights into their decision-making processes. (b) Depiction of the decision-making process by human experts for skin lesions, grounded in domain-specific knowledge and precise criteria, facilitates explainable diagnoses. (c) Overview of our Explicd framework: Domain knowledge is queried from LLMs or human experts across criteria axes. Explicd then aligns encoded visual concepts with textual knowledge anchors, facilitating the learning of visual concepts. The final diagnostic prediction is made based on the alignment scores between visual and textual concepts with a linear function.

size, or specific patterns regarding the disease symptom, staging, and outcome. In essence, human-expert decisions are grounded in a set of differential diagnostic criteria, enabling us to distinguish between various medical conditions with confidence and transparency. Developing explainable models that mirror the nuanced criteria defined by human knowledge is crucial to real-world AI clinical applications.

Incorporating human knowledge into the medical image assessment faces a long-standing challenge in the cross-modal data integration. The surge of vision-language models (VLMs) [24,13,20] has focused on aligning image-text pairs in a unified representation space, opening up opportunities for the joint understanding of vision and language tasks. However, major general VLMs [24] are not trained on medical data. They often fall short when dealing with medical image-text pairs due to a significant shift in data distribution [32]. Although efforts have been placed on VLMs [4,31] pretrained on biomedical data, their performance

still lags behind task-specific models in a broad range of medical image analysis tasks [29,8]. This is because they often struggle to align the associations between nuanced task-specific texts and the corresponding visual features. This inherent gap necessitates the development of better fine-grained alignment methods.

In this paper, we present **Explicd**, a simple yet effective framework for **Exp**lainable **l**anguage-**i**nformed **c**riteria-based **d**iagnosis. Typically, medical image classification benchmarks provide only the images and corresponding labels, largely omitting the detailed diagnostic information. Explicd addresses this gap by querying domain knowledge from either large language models (LLMs) like GPT-4 [1] or directly from human experts to formulate comprehensive diagnostic criteria, including key visual aspects such as color, shape, texture, or specific patterns associated with the classification task. We catalog the characteristics of each class based on these criteria axes. Within Explicd, these criteria are embedded as knowledge anchors using a VLM's text encoder. Moreover, we use a set of visual concept tokens to encode visual concepts from images along these criteria axes. An intermediate criteria anchor contrastive loss encourages high similarity between the encoded visual concepts and the corresponding positive knowledge anchors. Finally, a linear layer predicts the final diagnosis class by integrating the alignment scores from all criteria. Our contributions are as follows:

- We introduce Explicd as a simple yet effective framework for explainable language-informed, criteria-based diagnosis via vision-language models.
- Explicd utilizes domain knowledge from LLMs or human experts, defining diagnostic criteria for each class across specific concept axes.
- We propose a visual concept learning module alongside a criteria anchor contrastive loss to align fine-grained visual features with diagnostic criteria.
- Explicd demonstrates superior interpretability and performance compared to traditional black-box models on five public benchmarks.

2 Related Work.

Vision-language model. VLMs learn joint representations from image-text pairs through contrastive learning. Despite the strong generalizability demonstrated by pioneering works [24,13], applying VLMs to the biomedical domain is challenging due to the distribution shift and domain-specific vocabulary. To mitigate this, biomedical VLMs like BioViL [4] and BiomedCLIP [31] have been pretrained on biomedical data like radiology reports and PubMed articles and outperform general VLMs on some biomedical tasks, but still fall short compared to task-specific models [29]. In this study, we propose a knowledge-based, fine-grained alignment method that adapts biomedical VLMs to specific diagnostic criteria, bridging the performance gap with task-specific models in medical image classification.

Explainable model. Explainable AI models can be categorized into post-hoc and self-interpretable methods with a goal of making decision processes understandable [15]. Post-hoc methods (e.g., Grad-CAM [25]) analyze the trained

model to identify informative features, offering flexibility for pre-trained models, but they may not accurately reflect the model's true reasoning process [30]. On the other hand, self-interpretable methods design the explainability architecture directly inside the model. Concept bottleneck model (CBM) [19] is a representative work that predicts predefined concepts to enable transparent decision-making, but it requires time-consuming attribute annotation. Our approach aligns with self-interpretable models but leverages large language models (LLMs) to bypass the annotation of concepts. LaBo [30] is a state-of-the-art self-interpretable model using visual-language concept scores, but relies heavily on the quality of pretrained VLM alignment. Our method enhances explainability by specifying diagnostic criteria axes and proposing visual concept learning for better alignment, ensuring learned concepts are strongly related to human experts' diagnostic criteria.

3 Method

Fig. 1 (C) illustrates the proposed Explicd framework. In this section, we present details on how Explicd queries knowledge as diagnostic criteria, aligns visual features with these criteria, and makes explainable classifications.

3.1 Domain Knowledge Query & Diagnostic Criteria Formulation

Disease diagnosis usually centers around various criteria axes describing distinctive characteristics across clinical classes [22,28]. Drawing from inspiration, we first query domain knowledge from LLMs or consult human experts and formulate them into textual diagnosis criteria. Consider a set of training image-label pairs $D = \{(x, y)\}$, where x is the image and $y \in \mathcal{Y}$ is a label from a set of N classes. Specifically, we categorize the diagnosis criteria along K disentangled *criteria axes* specified by language depending on the task $\{C_i\}_{i=1}^{K}$. For instance, in the case of skin lesions, the criteria axes include `asymmetry, border, color, diameter, texture, pattern`. Subsequently, we query detailed knowledge on the typical characteristics for each class along each criteria axis $C_i = \{c_i^1, \ldots, c_i^{n_i}\}$, where $1 < n_i \leq N$ denotes the number of possible options within a particular criteria axis. Take the `color` of skin lesion as an example, potential options could range from 'a mixture of black brown and blue' for melanoma, 'uniform pigmentation of tan or brown' for melanocytic nevus, among others. Notably, the quantity of typical characteristics for each criteria axis, n_i, may be less than the number of classes N, as different classes might exhibit identical characteristics for certain criteria axes; e.g., various types of benign skin lesions could all present symmetry. Additionally, the ground truth label for each diagnostic criterion is recorded, i.e. for each criteria axis, the associated class and characteristic options are marked as positive, whereas all other combinations are considered negative.

3.2 Visual Concept Learning

After collecting the textual form diagnostic criteria, we aim to align the visual features with these textual human knowledge. In particular, we propose a lightweight visual concept learning module for aligning the fine-grained visual concepts and the nuanced textual criteria. Specifically, given a pretrained vision-language model with visual encoder $\mathcal{V}$ and text encoder $\mathcal{T}$, we first encode the queried diagnostic criteria into criteria anchor embeddings, $\{\mathbf{e}_i = \mathcal{T}(C_i)\}_{i=1}^{K}$, where $\mathbf{e}_i \in \mathcal{R}^{n_i \times d}$, and d is the dimension of the embedded token. These criteria embeddings act as a sparse representation of human knowledge, serving as anchors to facilitate the learning of visual concepts.

To capture visual concepts effectively, our visual concept learning module employs a set of K learnable visual concept tokens $\mathbf{p} \in \mathcal{R}^{K \times d}$, with each token designated to represent one of the K criteria axes. For a given image x and its feature map $\mathcal{V}(x)$, the concept encoding process is formalized as follows:

$$\hat{\mathbf{p}} = \text{cross-attention}(\mathbf{p}, \mathcal{V}(x), \mathcal{V}(x)), \tag{1}$$

where $\mathbf{p}$ is the query and $\mathcal{V}(x)$ serves as the key and value of the cross-attention layer. The visual concept tokens $\mathbf{p}$ interact with the image feature map, thereby encoding the relevant visual concepts associated with specific criteria axes into $\hat{\mathbf{p}} \in \mathcal{R}^{K \times d}$.

The learning of visual concepts is facilitated by a contrastive loss. For each criteria axis, the aggregated visual concept token $\hat{\mathbf{p}}_i$ is compared against all characteristic embeddings $\mathbf{e}_i$, calculating a similarity score. The criteria anchor contrastive loss is as follows:

$$\mathcal{L}^{i}_{anchor}(\hat{\mathbf{p}}_i, \mathbf{e}_i) = -\log \frac{\exp(sim(\hat{\mathbf{p}}_i, \mathbf{e}_i^{\text{positive}}/\tau)}{\sum_{j=1}^{n_i} \exp(sim(\hat{\mathbf{p}}_i, \mathbf{e}_i^{j})/\tau)} \tag{2}$$

where τ denotes the temperature parameter that adjusts the softness of the softmax distribution and we use dot product as the similarity function. The criteria anchor contrastive loss aims to increase the similarity between the encoded visual concept token $\hat{\mathbf{p}}_i$ and the positive criteria anchor embeddings while decreasing its similarity with the embeddings of negative characteristics, ensuring a more discriminative learning of visual concepts along each diagnostic criteria axis.

3.3 Explainable classification

The above knowledge anchor loss $\mathcal{L}_{anchor}$ enables the alignment of encoded visual concepts with the corresponding characteristic options along each diagnostic criteria axis. Intuitively, the similarity scores between the encoded visual concept token $\mathbf{p_i}$ and the diagnostic criteria anchor $\mathbf{e}_i$ indicate the model's assessments for each diagnostic criterion. Mirroring the approach of human experts, who make their final diagnosis on the evaluations across multiple criteria, we use a linear layer to make prediction of the final class by integrating the alignment scores from all K criteria axes.

$$\hat{y} = W(\text{concat}(sim(\hat{\mathbf{p}}_1, \mathbf{e}_1), \ldots, sim(\hat{\mathbf{p}}_K, \mathbf{e}_K)))^{\intercal}, \tag{3}$$

where concat(,) represents the concatenation operation and W is the weights in the linear layer that inherently reflect the significance of each diagnostic criterion's contribution towards the overall class prediction.

During the training phase, we optimize a joint objective that includes both the criteria anchor contrastive loss with cross-entropy loss for the final classification:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{ce}}(\hat{y}, y) + \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}^{i}_{anchor}(\hat{\mathbf{p}}_i, \mathbf{e}_i) \tag{4}$$

The embeddings of textual criteria anchors are precomputed and stored, ensuring that the training and inference overhead introduced by the additional components are negligible.

4 Experiments

4.1 Experimental Setup

We evaluate our method on five publicly available medical image classification benchmarks, which cover a diverse range of medical targets and modalities.
Dataset: ISIC2018 [27] contains 10,015 dermoscopic images with seven skin lesion categories for skin cancer classification. **NCT-CRC-HE**(NCT) [16] includes 100,000 patch-based histological images of human colorectal cancer for training and 7180 patches for validation, with nine tissue classes for classification. **IDRiD** [23] consists of 516 retinal fundus images annotated with 5 severity level grading of diabetic retinopathy. **BUSI** [3] dataset contains 780 ultrasound images of breast masses categorized into normal, benign, and malignant classes for breast cancer classification. **MIMIC-CXR** [14] contains 377,100 chest X-ray images. Cardiomegaly (CM) and Edema are used for binary classification.
Baselines. We compare our method with several baselines, including (1) **VLMs zero-shot**: We apply the general VLM CLIP [24] and biomedical VLMs BioViL [4] and BiomedCLIP [31] in a zero-shot setting for classification; (2) **Supervised black-box models**: We fine-tune ImageNet-pretrained ResNet50 [12] and ViT-Base [7] on the classification benchmarks; and (3) **LaBo**: a state-of-the-art explainable model with concept bottleneck.
Implementation Details. We prompt GPT-4 to query domain knowledge and diagnostic criteria. We use the official implementation and pretrained weights of CLIP ViT-Base, BioViL and BiomedCLIP. Our Explicd and LaBo are implemented based on BioViL-specialized for MIMIC-CXR dataset, while using BiomedCLIP for all other datasets. The fine-tuning of Explicd involves optimizing visual encoder, visual concept learning module and the final linear layer with AdamW optimizer, while keeping the text encoder fixed. All experiments are conducted using PyTorch with Nvidia A6000 GPUs.

4.2 Main Results

Our proposed Explicd model demonstrates superior performance compared to various baseline methods across five medical image classification benchmarks, as

Table 1: Performance comparison across five benchmarks. Balanced accuracy is reported for CM and edema in MIMIC-CXR due to class imbalance; accuracy is reported for the other datasets.

Setting	Model	ISIC2018	NCT	IDRiD	BUSI	CM	Edema
Zero-shot	CLIP	11.6	9.9	31.1	30.8	49.5	51.4
	BioViL	8.5	7.7	26.2	30.8	70.8	76.9
	BiomedCLIP	21.2	35.3	37.9	37.2	69.3	77.1
Black-box	ResNet50	82.6	93.4	53.4	84.6	79.7	77.4
	ViT-Base	89.0	94.4	57.3	88.5	79.2	80.9
Explainable	LaBo	80.9	90.2	48.4	75.8	73.5	74.2
	Explicd (ours)	**90.0**	**95.1**	**58.5**	**89.7**	**81.8**	**85.7**

shown in Table 1. The zero-shot performance of VLMs, including CLIP, BioViL, and BiomedCLIP, is generally poor, with CLIP performing close to random guessing across all datasets. This indicates that CLIP's visual-text alignment is not effective for complex medical diagnosis tasks as it is trained on general vision data. Although BioViL and BiomedCLIP perform much better on the MIMIC-CXR dataset (CM and Edema tasks), likely due to their pretraining data being largely based on chest X-ray radiology reports, their performance on other datasets remains much lower than supervised trained black-box models like ResNet50 and ViT-Base, suggesting their limited generalization ability.

Explicd effectively combines explainability with high-level classification performance, outperforming not only the explainable model LaBo but also black-box models across all datasets. LaBo's lower accuracy compared to the black-box models can be attributed to its reliance on well-aligned vision-language models, highlighting the challenge of maintaining high accuracy while providing strong explainability. In contrast, Explicd's superiority is due to the introduction of human knowledge and visual concept learning, which provides additional supervision for fine-grained alignment between visual features and diagnostic criteria. This strategy not only enhances fine explainability but also brings improvement in overall classification performance.

4.3 Diagnostic Interpretation

A distinguishing design of Explicd is its appealing ability to interpret its decision-making process. Fig. 2 (a) shows the alignment scores measured with cosine similarity between the encoded visual concept tokens and the embeddings of diagnostic criteria for skin lesions. The width of the lines indicates the strength of similarity, with larger widths representing higher similarity scores. We can see that Explicd can accurately predict the characteristics of each criteria axis, such as the presence of asymmetry, border irregularity, color variegation, and large diameter, which are key features in the diagnostic criteria we queried for melanoma diagnosis. The high similarity scores between the visual concepts and the corresponding diagnostic criteria demonstrate that Explicd has learned to identify and align these important visual features, leading to a correct final diagnosis of melanoma.

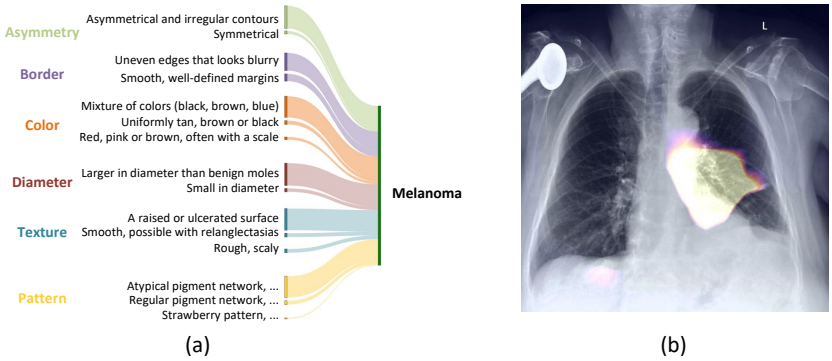

Fig. 2: (a) Alignment scores measured using cosine similarity between the encoded visual concept tokens and diagnostic criteria along each axis for skin lesion classification. The width of the lines represents the strength of similarity, with wider lines indicating higher scores. (b) Heatmap visualization of the encoded visual concept tokens overlaid on the image feature maps for a case of cardiomegaly. Brighter regions indicate higher similarity scores, suggesting a stronger focus on these areas by the model.

Furthermore, we visualize the heatmap of the average visual concept tokens with the image feature map of cardiomegaly in a chest X-ray image in Fig. 2 (b). The brighter regions indicate higher similarity scores, suggesting that the model is focusing on these areas when making its prediction. Cardiomegaly is a medical condition characterized by an enlarged heart. In the heatmap, we can observe that Explicd correctly focuses its attention on the heart area, indicating that Explicd has learned to align human knowledge regarding the key visual features of cardiomegaly with the relevant visual concepts in the X-ray image. By providing the alignment scores on criteria and highlighting the most important regions contributing to its prediction, Explicd provides a transparent and interpretable decision-making process that can be easily understood and verified by medical experts.

5 Discussion

To address the lack of transparency and interpretability in current deep learning models for medical image analysis, we proposed Explicd, a comprehensive framework that integrates diagnostic criteria queried from LLMs, aligning visual concepts towards explainable classification. Explicd offers a novel means to understanding diseases along human-understandable criteria axes. Our extensive experiments highlight Explicd's superior performance over both traditional black-box approaches and existing explainable models, setting a new standard in both accuracy and interpretability. The clarity of Explicd's decision-making process promises to bolster trust and facilitate the integration of AI in clini-

cal diagnostics. Moving forward, we aim to expand the incorporation of broad human knowledge within our diagnostic criteria and refine the hierarchical representations of visual concepts, allowing for a more nuanced exploration of disease diagnosis and management.

Acknowledgments. This research has been partially funded by research grants to D. Metaxas through NSF: 2310966, 2235405, 2212301, 2003874, and FA9550-23-1-0417 and NIH 2R01HL127661.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D.S., Karthikesalingam, A., King, D., Ashrafian, H., Darzi, A.: Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ digital medicine **4**(1), 65 (2021)
3. Al-Dhabyani, W., Gomaa, M., Khaled, H., Aly, F.: Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. Int. J. Adv. Comput. Sci. Appl **10**(5), 1–11 (2019)
4. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
5. Cai, L., Gao, J., Zhao, D.: A review of the application of deep learning in medical image classification and segmentation. Annals of translational medicine **8**(11) (2020)
6. Cutillo, C.M., Sharma, K.R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K.D., in Healthcare Workshop Working Group Beck Tyler 1 Collier Elaine 1 Colvis Christine 1 Gersing Kenneth 1 Gordon Valery 1 Jensen Roxanne 8 Shabestari Behrouz 9 Southall Noel 1, M.: Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. NPJ digital medicine **3**(1), 47 (2020)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Gao, Y.: Training like a medical resident: Context-prior learning toward universal medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11194–11204 (2024)
9. Gao, Y., Huang, R., Chen, M., Wang, Z., Deng, J., Chen, Y., Yang, Y., Zhang, J., Tao, C., Li, H.: Focusnet: imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 829–838. Springer (2019)

10. Gao, Y., Huang, R., Yang, Y., Zhang, J., Shao, K., Tao, C., Chen, Y., Metaxas, D.N., Li, H., Chen, M.: Focusnetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck ct images. Medical Image Analysis **67**, 101831 (2021)
11. Gao, Y., Zhou, M., Metaxas, D.N.: Utnet: a hybrid transformer architecture for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 61–71. Springer (2021)
12. 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)
13. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916. PMLR (2021)
14. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
15. Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C., Medya, S.: A survey on explainability of graph neural networks. arXiv preprint arXiv:2306.01958 (2023)
16. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo10 **5281** (2018)
17. Khanna, N.N., Maindarkar, M.A., Viswanathan, V., Fernandes, J.F.E., Paul, S., Bhagawati, M., Ahluwalia, P., Ruzsa, Z., Sharma, A., Kolluri, R., et al.: Economics of artificial intelligence in healthcare: diagnosis vs. treatment. In: Healthcare. vol. 10, p. 2493. MDPI (2022)
18. Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., Ganslandt, T.: Transfer learning for medical image classification: a literature review. BMC medical imaging **22**(1), 69 (2022)
19. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: ICML. pp. 5338–5348. PMLR (2020)
20. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
21. Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D.D., Chen, M.: Medical image classification with convolutional neural network. In: 2014 13th international conference on control automation robotics & vision (ICARCV). pp. 844–848. IEEE (2014)
22. Liberman, L., Menell, J.H.: Breast imaging reporting and data system (bi-rads). Radiologic Clinics **40**(3), 409–430 (2002)
23. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
24. 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: ICML. pp. 8748–8763. PMLR (2021)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE ICCV. pp. 618–626 (2017)
26. Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: Information Processing in Medical

Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24. pp. 588–599. Springer (2015)
27. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
28. Turkbey, B., Rosenkrantz, A.B., Haider, M.A., Padhani, A.R., Villeirs, G., Macura, K.J., Tempany, C.M., Choyke, P.L., Cornud, F., Margolis, D.J., et al.: Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. European urology **76**(3), 340–351 (2019)
29. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Vision-language modelling for radiological imaging and reports in the low data regime. arXiv preprint arXiv:2303.17644 (2023)
30. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19187–19197 (2023)
31. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 **2**(3), 6 (2023)
32. Zhang, Y., Gao, J., Tan, Z., Zhou, L., Ding, K., Zhou, M., Zhang, S., Wang, D.: Data-centric foundation models in computational healthcare: A survey. arXiv preprint arXiv:2401.02458 (2024)