CoPA: Hierarchical Concept Prompting and Aggregating Network for Explainable Diagnosis

Yiheng Dong[1*], Yi Lin[2*], and Xin Yang[1(✉)]

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China
xinyang2014@hust.edu.cn
[2] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

Abstract. The transparency of deep learning models is essential for clinical diagnostics. Concept Bottleneck Model provides clear decision-making processes for diagnosis by transforming the latent space of black-box models into human-understandable concepts. However, concept-based methods still face challenges in concept capture capabilities. These methods often rely on encode features solely from the final layer, neglecting shallow and multiscale features, and lack effective guidance in concept encoding, hindering fine-grained concept extraction. To address these issues, we introduce Concept Prompting and Aggregating (CoPA), a novel framework designed to capture multilayer concepts under prompt guidance. This framework utilizes the Concept-aware Embedding Generator (CEG) to extract concept representations from each layer of the visual encoder. Simultaneously, these representations serve as prompts for Concept Prompt Tuning (CPT), steering the model towards amplifying critical concept-related visual cues. Visual representations from each layer are aggregated to align with textual concept representations. With the proposed method, valuable concept-wise information in the images is captured and utilized effectively, thus improving the performance of concept and disease prediction. Extensive experimental results demonstrate that CoPA outperforms state-of-the-art methods on three public datasets. Code is available at https://github.com/yihengd/CoPA.

Keywords: Explainable Diagnosis · Concept Bottleneck Model · Prompt Tuning.

1 Introduction

Rapid advancements in deep learning have led to remarkable progress in medical image analysis [15,1,22], yet ensuring model interpretability remains a critical challenge in clinical practice [23]. Traditional post-hoc interpretability methods, such as CAM [29] and Grad-CAM [21], which provide visualizations of model decisions, often fail to meet the high precision and reliability demands of

* These authors contributed equally to this work.

medical scenarios [14,20]. To address this limitation, Concept Bottleneck Model (CBM) [13] has emerged as a promising solution. CBM maps input images to a predefined set of human-interpretable concepts, serving as a verifiable "bridge" between raw inputs and final predictions, making the decision-making process interpretable. These concepts encompass a spectrum of features, ranging from low-level attributes like color and shape to high-level semantic features, such as ulceration, providing deep insights into the model's decision-making process.

However, existing CBM implementations [26,5,6,2,17] often exhibit limited effectiveness in capturing concepts, as they typically rely on final-layer features of the image encoder (ResNet/ViT) for concept alignment. As noted by [18,24], although these deep representations often tend to capture high-level global semantics, they inevitably overlook critical low-level and local visual patterns. This representation deficiency leads to the inadequate encoding of concepts that require shallow or multiscale analysis (e.g., dots and globules), ultimately impacting concept alignment and disease diagnosis.

Recent integration of Vision-Language Models (VLMs) with CBMs has revealed new opportunities through their pre-trained cross-modal alignment capabilities [26,6,2,25]. However, due to the scarcity of annotated medical datasets, VLMs face challenges in capturing fine-grained concept semantics, particularly for recognizing complex pathological patterns. Furthermore, the issue of model forgetting [27] warrants attention. After fine-tuning on downstream tasks, models may significantly forget previously encoded specialized medical knowledge.

To address these challenges, we propose **Co**ncept **P**rompting and **A**ggregating network (CoPA) aiming at enabling fine-grained and multiscale feature capture and differentiation for concepts. Specifically, Concept-aware Embedding Generator (CEG) is proposed to distill highly concentrated concept-aware feature representations, which are aggregated by a selector to form the final visual concept representation. Subsequently, we introduce Concept Prompt Tuning (CPT), where the outputs of CEG serve as inputs to the next transformer layer with the backbone frozen, guiding the concept prompt to progressively concentrate more on concept-related features. Contrastive learning is then utilized to align visual and textual concept representations. Finally, a gating network is employed to weigh and combine the aligned concept representations for disease prediction.

The main contributions of our work are as follows: 1) We present Concept Prompting and Aggregating (CoPA), a novel explainable network adept at capturing multiscale and fine-grained visual concept representations. 2) We design Concept-aware Embedding Generator (CEG), which extracts highly concentrated concept-related visual representations from each layer of the visual encoder, facilitating multilayer feature aggregation. 3) We design Concept Prompt Tuning (CPT) technique to guide the model's focus on target visual concepts and mitigate the issue of knowledge forgetting. 4) Extensive experimental results demonstrate that our CoPA outperforms state-of-the-art methods, highlighting the efficacy of each component in our framework.

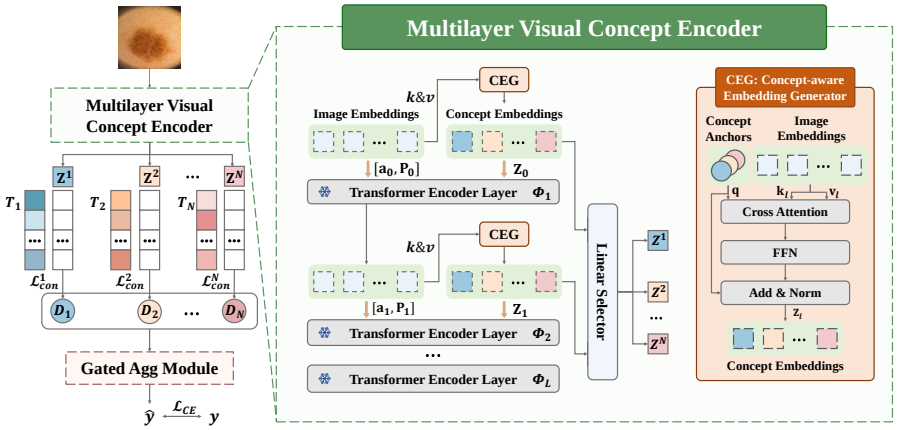

Fig. 1. The overall pipeline of CoPA, which consists of a multi-layer visual concept encoder, a concept alignment bottleneck layer, and a gated aggregation module.

2 Method

2.1 Overall Pipeline

The architecture of the proposed model is shown in Fig. 1. Given a data represented by triplets $\mathcal{D} = \{(x, c, y)\}$, where x denotes the input image, y represents the disease label and $c = \{c_1, c_2...c_N\}$ signifies a set of concept labels with the number of N. Moreover, each concept c_i belongs to a candidate set $\mathcal{C}_i = \{c_i^1, c_i^2, ..., c_i^{k_i}\}$ (e.g., for the concept "Pigment Network", $\mathcal{C}_i$ = {"atypical", "typical"}), where k_i indicates the number of elements in $\mathcal{C}_i$ for the i^{th} concept.

We start with passing the image x through Multilayer Visual Concept Encoder to generate concept-aware visual embeddings. In parallel, text embeddings for each concept candidate set $\mathcal{C}_i$ are generated using a frozen text encoder. Afterwards, cross-modal alignment is achieved through contrastive learning between two modalities. Finally, the aligned representations undergo adaptive fusion and generate disease predictions.

2.2 Multilayer Visual Concept Learning

Visual concept representation extraction comprises two critical components operating at each visual encoder layer. Unlike some prior methods that predict all concepts using a single feature map, our framework leverages concept-aware visual embeddings from Concept-aware Embedding Generator (CEG) to effectively encode the relevant visual features associated with specific concepts. However, the model's capability to localize concept-aware visual information remains suboptimal. We attribute this to insufficient semantic guidance during encoding and catastrophic forgetting during full-parameter fine-tuning [27]. Therefore, Concept Prompt Tuning (CPT) is proposed to inject semantic guidance into feature

extraction. Hierarchically aggregated layer-wise representations yield a unified visual concept embedding for cross-modal alignment.

Concept-aware Embedding Generator (CEG). Our CEG takes two sides of input: a set of learnable concept anchors and a token-wise feature map from the visual encoder layer. Each anchor $\mathbf{q}_i \in \mathbb{R}^d, 1 \leq i \leq N$, where N denotes the length of concept set c, is associated with a specific concept and serves as a query vector to retrieve essential information from visual features. For a given input from the l^{th} encoder layer, we consider image embedding as both key and value vectors $\mathbf{k}_l, \mathbf{v}_l \in \mathbb{R}^{m \times d}$, where m stands for their total numbers. The concept-aware embeddings $\mathbf{z}_l^i \in \mathbb{R}^d$ are computed as:

$$\hat{\mathbf{z}}_l^i = \text{Softmax}(\frac{\mathbf{q}_i \mathbf{k}_l^\top}{\sqrt{d_k}})\mathbf{v}_l, 1 \leq i \leq N, \tag{1}$$

$$\mathbf{z}_l^i = LN(FFN(\hat{\mathbf{z}}_l^i) + \mathbf{q}_i), 1 \leq i \leq N, \tag{2}$$

where $FFN(\cdot)$ denotes feed forward network and $LN(\cdot)$ denotes layer normalization. By performing this operation across all encoder layers, we obtain L embeddings, which are then aggregated by a linear selector to produce multi-layer embeddings $\mathbf{Z}^i$. In addition, for the l^{th} layer, $\mathbf{z}_l^i$ also serves as input parallel to the image embeddings, as will be discussed below.

Concept prompt tuning (CPT). Motivated by visual prompt tuning (VPT) [10], we introduce concept prompt tuning. This design preserves the rich representations of pre-trained models while substantially addressing the performance degradation from excessive parameter tuning. Moreover, CPT supports concept-specific prompt design, thereby enabling self-attention mechanisms within each layer to progressively amplify target visual concepts, which significantly refines feature representations in a task-driven manner.

To be specific, let patch embeddings $\mathbf{P}_l = \left\{ \mathbf{p}_l^{k_p} \in \mathbb{R}^d \mid k_p \in \mathbb{N}, 1 \leq k_p \leq N_p \right\}$ of length N_p and class token embedding $\mathbf{a}_l \in \mathbb{R}^d$ denoted as input to l^{th} layer Φ_l of the original image encoder $\Phi(\cdot;\theta)$. We introduce concept-aware visual embeddings $\mathbf{Z}_l = \left\{ \mathbf{z}_l^i \in \mathbb{R}^d \mid i \in \mathbb{N}, 1 \leq i \leq N \right\}$, concatenated with $\mathbf{a}_l$ and $\mathbf{P}_l$, while keeping all backbone weights θ frozen. The entire process can be formulated as:

$$[\mathbf{a}_l, \_, \mathbf{P}_l] = \Phi_l\left([\mathbf{a}_{l-1}, \mathbf{Z}_{l-1}, \mathbf{P}_{l-1}]\right), l = 1, 2, \ldots, L. \tag{3}$$

2.3 Explainable Diagnose

For each concept, its candidate set $\mathcal{C}_i$ is formatted into structured text templates (e.g. "*This is a dermoscopic image, the* {*concept title*} *of the lesion is* $\{c_i^j\}$") and encoded by a pre-trained text encoder into embeddings $T_i = [t_i^1, ..., t_i^{k_i}] \in \mathbb{R}^{k_i \times d}$, where k_i is the length of $\mathcal{C}_i$. During the concept alignment phase, contrastive learning is applied between visual and textual concept features to maximize their semantic consistency through alignment loss formulated as:

$$\mathcal{L}_{con} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(\mathbf{Z}^i, t_i^*/\tau))}{\sum_{j=1}^{k_i} \exp(sim(\mathbf{Z}^i, t_i^j/\tau))}, \tag{4}$$

where t_i^* is the true label of concept c_i, τ is a learnable temperature parameter and $sim(\cdot,\cdot)$ represents cosine similarity. Afterwards, textual embeddings are fused into D_i within the concept candidate set $\mathcal{C}_i$ according to their normalized similarity alignment scores for disease diagnosis.

Emulating clinical decision-making workflows where experts synthesize multi-criteria assessments to derive conclusions and modified by [13], gated aggregation module is employed to adaptively combine concept-aligned representations from all learned concepts through learnable weights α that quantify each concept's diagnostic relevance:

$$\hat{y} = FC(\sum_{i=1}^{N} \alpha_i \cdot D_i), \tag{5}$$

where FC denotes full-connection layer to make the final decision and D_i denotes the i^{th} fused textual concept representation.

During the training phase, the optimization process simultaneously minimizes a composite loss function comprising concept alignment loss and diagnostic cross-entropy loss to supervise both concept and disease classification:

$$\mathcal{L} = \lambda \mathcal{L}_{con} + (1-\lambda)\mathcal{L}_{CE}(\hat{y}, y), \tag{6}$$

where $\lambda \in [0,1]$ is a hyperparameter controlling the relative importance of concept and disease accuracy.

3 Experiment

3.1 Experiment Setup

Datasets. **PH^2** [16] comprises 200 dermoscopic images, with annotations for five morphological features. By merging subcategories encompassing common and atypical nevi classes, the final dataset consists of 160 nevus and 40 melanoma cases. **Derm7pt** [12] consists of 1,011 dermoscopic image cases, with annotations based on the clinical 7-point checklist. Following Bie et al. [2], we retain all seven clinical indicators with 575 nevi and 252 melanoma cases. **SkinCon** [4] contains 3,690 clinical images from Fitzpatrick 17k [7]. In this study, 22 high-frequency clinical features, each with over 50 annotation instances, are selected. The disease categories include non-neoplastic, benign, and malignant. All datasets are randomly divided into training, validation, and test sets in a 70%:15%:15% ratio.

Implementation Details. Our framework initializes both image and text encoders using BiomedCLIP [28] pre-trained weights and settings. Optimization is performed using Adam with a learning rate η=1e-5. The loss weighting coefficient λ controlling concept-task balance is set to 0.5 via cross-validation. All experiments are implemented in PyTorch and executed on NVIDIA GeForce RTX 3090 GPUs, with results averaged over three random seeds.

Table 1. Quantitative comparison results in **disease prediction** in terms of area Under Curve (AUC), accuracy and F1-score. Results are in percentage(%).

Method	**PH^2**			**Derm7pt**			**SkinCon**		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
PCBM [26]	$78.3_{1.2}$	$89.3_{1.9}$	$81.5_{2.6}$	$73.0_{2.2}$	$77.0_{1.4}$	$71.0_{1.2}$	$68.9_{1.6}$	$71.0_{1.1}$	$70.5_{0.8}$
PCBM-h [26]	$92.3_{1.5}$	$90.7_{1.9}$	$83.3_{2.6}$	$83.3_{1.1}$	$79.9_{0.9}$	$74.5_{1.4}$	$69.5_{1.7}$	$72.3_{1.4}$	$72.3_{1.3}$
CBE [17]	$97.5_{0.0}$	$96.0_{0.0}$	$93.9_{0.0}$	$76.6_{0.4}$	$83.8_{0.3}$	$78.1_{0.4}$	$72.8_{1.2}$	$73.8_{1.1}$	$73.6_{1.3}$
Explicd [6]	$95.4_{2.4}$	$94.4_{2.3}$	$92.8_{3.6}$	$87.5_{3.2}$	$81.0_{3.2}$	$80.5_{3.5}$	$74.0_{0.9}$	$73.1_{0.3}$	$72.0_{0.7}$
MICA [2]	$98.2_{1.4}$	$98.7_{1.9}$	$95.3_{1.2}$	$85.6_{1.1}$	$83.9_{1.0}$	$79.4_{1.3}$	$75.9_{1.1}$	$75.6_{1.1}$	$75.4_{1.2}$
Our work	$\mathbf{98.3_{1.6}}$	$\mathbf{98.9_{2.2}}$	$\mathbf{98.8_{2.3}}$	$\mathbf{92.1_{1.2}}$	$\mathbf{86.0_{0.5}}$	$\mathbf{85.8_{0.9}}$	$\mathbf{77.5_{0.6}}$	$\mathbf{76.3_{0.8}}$	$\mathbf{75.7_{0.8}}$

3.2 Experiment Result

Comparison with existing methods. To evaluate the efficacy of our approach, we compare it with existing concept-based methods on the aforementioned datasets. Methods for comparison include: PCBM [26], CBE [17], Explicd [6], and MICA [2]. Table 1 and 2 summarize results for concept alignment and disease classification, respectively. Our framework achieves state-of-the-art performance across most metrics, especially on Derm7pt, where CoPA's disease prediction accuracy exceeds the second-best by 2.1%, and on PH^2 with a concept prediction accuracy 2.6% higher than the second-best approach.

Ablation study. We conduct various ablation studies on PH^2 and Derm7pt to investigate the influence of different modules and settings. Table 3 quantifies the contribution of individual components to overall performance. Specifically, our ablation analysis show that all designed components contribute positively, including 1) MultiLayer Aggregation strategy (MLA), which hierarchically aggregates features across encoder layers to capture multiscale features; 2) Concept Prompt Tuning (CPT), designed to highlight the concept-specific feature; and 3) Frozen Vision Backbone (FVB) that address knowledge forgetting caused by parameter-efficient fine-tuning. Table 3 shows that our approach, including all three components, achieves optimal performance.

Table 2. Quantitative comparison results in **concept prediction** in terms of AUC, accuracy and F1-score. Results are in percentage(%).

Method	PH^2			Derm7pt			SkinCon		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
CBE [17]	81.3	71.6	70.0	72.2	74.1	71.0	79.3	89.0	62.1
Explicd [6]	88.8	79.6	76.4	85.7	73.0	71.9	76.0	93.1	65.5
MICA [2]	83.6	75.2	68.4	78.6	76.0	72.4	**82.6**	91.7	63.8
Our work	**89.0**	**82.2**	**80.6**	**87.0**	**77.1**	**76.6**	81.7	**93.6**	**70.4**

Table 3. Ablation study of CoPA on PH^2 and Derm7pt. MLA, CPT and FVB represent Multi-Layer Aggregation, Concept Prompt Tuning and Freezing Vision Backbone.

MLA	CPT	FVB	PH^2				Derm7pt			
			Label		Concept		Label		Concept	
			ACC	F1	ACC	F1	ACC	F1	ACC	F1
✗	✗	✗	93.3	93.9	79.6	76.4	81.7	81.2	73.0	71.9
✔	✗	✗	96.7	96.5	80.9	78.1	85.0	84.2	73.1	72.2
✗	✔	✗	96.7	95.2	80.0	76.6	84.5	85.0	73.6	73.1
✗	✔	✔	96.7	96.7	80.5	81.7	85.0	85.2	76.0	75.2
✔	✔	✗	97.8	96.4	81.1	79.1	85.2	85.5	74.0	73.2
✔	✔	✔	**98.9**	**98.8**	**82.2**	**80.6**	**86.0**	**85.8**	**77.1**	**76.6**

3.3 Interpretability Analysis

Inspired by previous work [8,9,19], we analyze the interpretability of our model from the following three aspects: *faithfulness*, *understandability*, and *plausibility*.

Faithfulness. *Faithfulness* refers to the extent to which explanations accurately reflect the internal decision-making process of the model [8,19]. In this study, we evaluate model's faithfulness through manual intervention during inference on Derm7pt. Specifically, as shown in Fig. 2, for positive intervention, we set 1-2 incorrectly predicted concepts' ground-truth confidence to 1, while for negative intervention, we adjust 1-2 correctly predicted concepts' ground-truth confidence to 0, to observe the outcome changes of the disease prediction. Notably, other confidence of adjusted concepts is recalculated by softmax function to ensure the probability sum equals 1. As shown in Table 4, positive interventions resulted in accuracy increase of 0.5% (single-concept) and 1.1% (two-concept), whereas negative interventions led to accuracy reductions by 2.4% and 3.2%, respectively, showing model's heavy reliance on concept predictions and affirming the faithfulness of the explanations.

Understandability & Plausibility. *Understandability* refers that the explanation's context should be readily comprehensible to users, eliminating the necessity for technical expertise [11], while *plausibility* refers to the extent to which explanations align with domain-specific human reasoning and appear credible [3]. Fig. 3 shows the examples of explanations in detail. In Fig. 3(a), we visualize concept-associated regions and their prediction confidence scores, providing users with a foundation to assess the acceptability of concept alignment. Fig. 3(b) illustrates the process of concept prediction and disease diagnosis for a data sample, including visual concept heatmaps, concept confidence scores, gated network

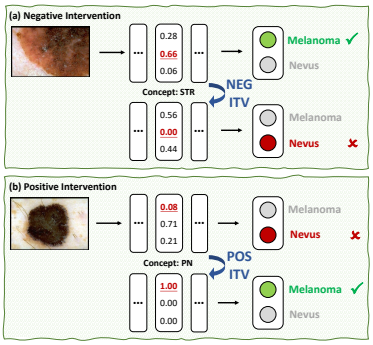

Fig. 2. Intervention Examples. ITV: Intervention

	ITV nums	ACC	Improve
ITV-Free	0	86	-
Neg ITV	1	83.6	-2.4
	2	82.8	-3.2
Pos ITV	1	86.5	0.5
	2	87.1	1.1

Table 4. Accuracy changes of test-time concept intervention on Derm7pt.

weights, and diagnostic confidence. This workflow provides users with a transparent and traceable insight into the entire prediction process, ensuring the interpretability of diagnostic decisions.

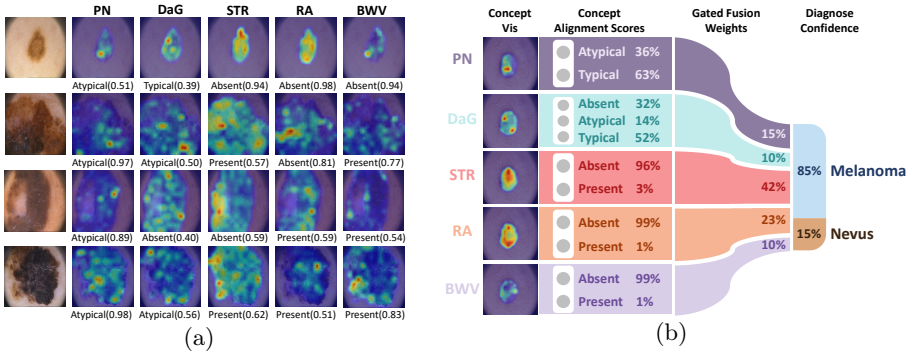

Fig. 3. Illustration of understandability and plausibility, where PN, DaG, STR, RA, BWV stand for "Pigment Network", "Dots and Globules", "Streaks", "Regression Area", "Blue-Whitish Veil", respectively. (a) Heatmap visualization of concern areas for each concept. (b) The entire process of the prediction, including concept visualization, concept alignment scores, gated fusion mechanism weights, and diagnose confidence.

4 Conclusion

In this paper, we propose CoPA, a multilayer concept prompting and aggregation framework for interpretable disease diagnosis. Within the Concept-aware

Embedding Generator (CEG) of this framework, concept anchors is employed to query multi-scale visual features, generating densely concentrated concept representations that are hierarchically aggregated. Furthermore, to preserve the vast knowledge from the pre-trained vision-language model while enabling discriminative fine-tuning, we introduce Concept Prompt Tuning (CPT), which utilizes concept-aware representations as task-oriented visual prompts, guiding the model to focus on concept-relevant features. Experiments on three datasets demonstrate the exceptional performance and interpretability of our method.

Acknowledgments. This work was supported in part by the Natural Science Foundation of China under Grant 62472184, and in part by the Fundamental Research Funds for the Central Universities.

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

References

1. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3478–3488 (2021)
2. Bie, Y., Luo, L., Chen, H.: Mica: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 837–845 (2024)
3. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
4. Daneshjou, R., Yuksekgonul, M., Cai, Z.R., Novoa, R., Zou, J.Y.: Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. Advances in Neural Information Processing Systems **35**, 18157–18167 (2022)
5. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., et al.: Concept embedding models: Beyond the accuracy-explainability trade-off. Advances in Neural Information Processing Systems **35**, 21400–21413 (2022)
6. Gao, Y., Gu, D., Zhou, M., Metaxas, D.: Aligning human knowledge with visual concepts towards explainable medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2024)
7. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1820–1828 (2021)
8. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
9. Hsiao, J.H.w., Ngai, H.H.T., Qiu, L., Yang, Y., Cao, C.C.: Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI). arXiv preprint arXiv:2108.01737 (2021)

10. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
11. Jin, W., Li, X., Fatehi, M., Hamarneh, G.: Guidelines and evaluation of clinical explainable ai in medical image analysis. Medical image analysis **84**, 102684 (2023)
12. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics **23**(2), 538–546 (2018)
13. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
14. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: Unjustified counterfactual explanations. arXiv preprint arXiv:1907.09294 (2019)
15. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
16. Mendonça, T., Celebi, M., Mendonca, T., Marques, J.: Ph2: A public database for the analysis of dermoscopic images. Dermoscopy image analysis **2** (2015)
17. Patrício, C., Neves, J.C., Teixeira, L.F.: Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3799–3808 (2023)
18. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in neural information processing systems **34**, 12116–12128 (2021)
19. Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., Scotton, P.: Attention-based interpretability with concept transformers. In: International conference on learning representations (2021)
20. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. International journal of computer vision **128**, 336–359 (2020)
22. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical image analysis **88**, 102802 (2023)
23. Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis **79**, 102470 (2022)
24. Xu, Q., Wang, J., Jiang, B., Luo, B.: Fine-grained visual classification via internal ensemble learning transformer. IEEE Transactions on Multimedia **25**, 9015–9028 (2023)
25. 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)
26. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)

27. Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y.J., Ma, Y.: Investigating the catastrophic forgetting in multimodal large language models. In: NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
29. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)