Concept-Attention Whitening for Interpretable Skin Lesion Diagnosis

Junlin Hou[1], Jilan Xu[2], and Hao Chen[1,3(✉)]

[1] The Hong Kong University of Science and Technology, Hong Kong, China
[2] Fudan University, Shanghai, China
[3] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China

Abstract. The black-box nature of deep learning models has raised concerns about their interpretability for successful deployment in real-world clinical applications. To address the concerns, eXplainable Artificial Intelligence (XAI) aims to provide clear and understandable explanations of the decision-making process. In the medical domain, concepts such as attributes of lesions or abnormalities serve as key evidence for deriving diagnostic results. Existing concept-based models mainly depend on concepts that appear independently and require fine-grained concept annotations such as bounding boxes. However, a medical image usually contains multiple concepts, and the fine-grained concept annotations are difficult to acquire. In this paper, we aim to interpret representations in deep neural networks by aligning the axes of the latent space with known concepts of interest. We propose a novel Concept-Attention Whitening (CAW) framework for interpretable skin lesion diagnosis. CAW is comprised of a disease diagnosis branch and a concept alignment branch. In the former branch, we train a convolutional neural network (CNN) with an inserted CAW layer to perform skin lesion diagnosis. The CAW layer decorrelates features and aligns image features to conceptual meanings via an orthogonal matrix. In the latter branch, the orthogonal matrix is calculated under the guidance of the concept attention mask. We particularly introduce a weakly-supervised concept mask generator that only leverages coarse concept labels for filtering local regions that are relevant to certain concepts, improving the optimization of the orthogonal matrix. Extensive experiments on two public skin lesion diagnosis datasets demonstrated that CAW not only enhanced interpretability but also maintained a state-of-the-art diagnostic performance.

Keywords: Explainable AI · Concept Attention · Skin Lesion Diagnosis

1 Introduction

Deep learning has achieved significant advancements in medical image analysis. However, the black-box nature of deep learning greatly hinders its practical deployment and application [14, 7]. The networks usually output predictions without providing any explanation, resulting in a lack of interpretability. Therefore,

there is an urgent need to develop eXplainable AI (XAI) techniques that can enhance the transparency and understandability of the decision-making process.

Recently, there has been an increasing consensus that XAI should incorporate explanations based on *concepts* [20]. In the medical domain, concepts can be defined as high-level attributes of lesions or abnormalities, serving as evidence for deriving diagnostic results. For instance, *blue whitish veil*, *atypical pigmentation network*, and *irregular streaks* can be important concepts for diagnosing melanoma skin disease [12]. Given the concept annotations, current ante-hoc concept-based models mainly fall into two categories. The first is joint training for the target task and concept representation learning [17, 6, 16]. For example, Concept Bottleneck Model (CBM) [17] first predicted concepts and then made final predictions, but its generalization capability is much lower than standard end-to-end models. The second is to inject concepts from an external concept dataset to train deep neural networks by modifying a middle layer to represent concepts [3, 25, 21, 13]. Chen *et al.* [3] proposed a concept whitening layer, consisting of whitening and orthogonal transformation to align the concepts with the axes. These methods achieve good interpretability on natural images due to two primary reasons. First, a majority of images contain only one concept that is highly related to the category (*e.g.*, airplane→airfield). Second, they have fine-grained concept annotations (*e.g.*, bounding boxes) to crop concept regions from raw images. In this way, a set of most representative images that depict the concept can be collected to generate precise concept features.

However, medical images present a more complex challenge compared to natural images. A medical image usually contains multiple concepts, such as different types of lesions or abnormalities, that need to be considered for accurate classification. Moreover, obtaining fine-grained concept annotations such as masks or bounding boxes for medical images is time-consuming and laborious. The available concept information is often limited to coarse, image-level concept labels. To address the above issues, we propose a novel method named Concept-Attention Whitening (CAW) to enhance the representation interpretability of skin lesion diagnosis, where the axes of latent space are aligned with specific concepts. The main contributions of our method are three folds: (1) We establish a unique dual-branch optimization framework. In the disease diagnosis branch, given a disease dataset, we train a convolutional neural network (CNN) to perform disease classification with a novel CAW layer inserted. The CAW layer aims to decorrelate features by whitening transformation and assign conceptual meanings to specified dimensions via an orthogonal matrix. (2) In the concept alignment branch, we use a concept dataset to calculate the orthogonal matrix based on concept features. As an image may contain multiple concepts, we particularly introduce a weakly-supervised concept mask generator to produce concept-attentive masks by only using the concept labels. The concept mask highlights the most relevant local regions regarding a certain concept. In this way, we can obtain representative concept features and calculate an accurate orthogonal matrix by solving an optimization problem. (3) Extensive experiments were conducted on two skin lesion diagnosis datasets with concept annotations.

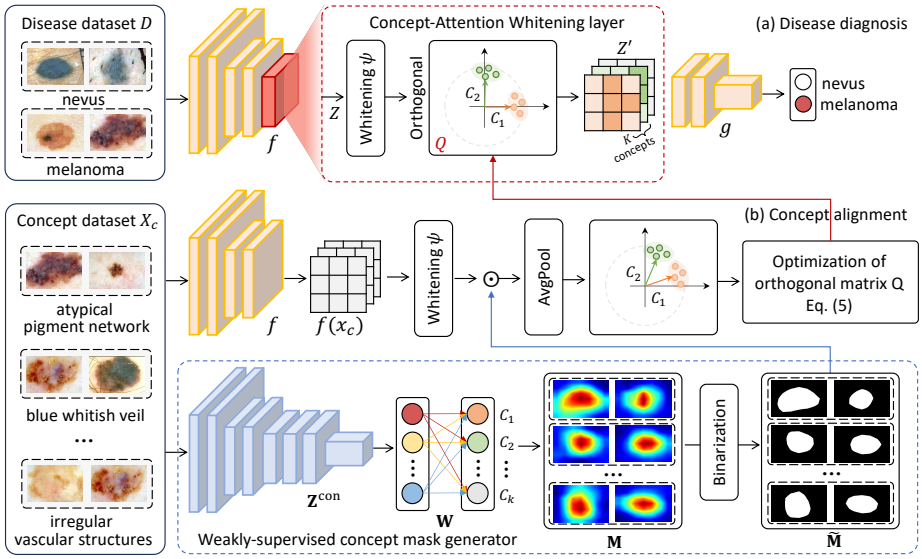

Fig. 1. An overall framework of the proposed Concept-Attention Whitening (CAW), including (a) a disease diagnosis branch and (b) a concept alignment branch.

Compared with existing state-of-the-art methods, our proposed method not only enhanced interpretability but also maintained a high diagnostic performance.

2 Methodology

Our overall objective is to train an interpretable disease diagnosis model that satisfies: (1) high disease diagnosis performance; and (2) the latent image features produced by the model are aligned to a pre-defined set of concepts. To achieve this goal, we propose a novel Concept-Attention Whitening (CAW) framework for interpretable skin lesion diagnosis using clinical concepts, as illustrated in Fig. 1. Specifically, a Concept-Attention Whitening layer is inserted into the encoder network to disentangle concepts and align the latent image features with known concepts of interest. For network training, we adopt a dual-branch architecture, comprised of disease diagnosis and weakly-supervised concept alignment.

2.1 Disease Diagnosis Branch

Given the disease dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ where y_i is the disease label of sample x_i, we train a CNN (*e.g.*, ResNet-50) to classify the skin disease. We replace the Batch Normalization (BN) layer with our Concept-Attention Whitening (CAW) layer to produce intepretable representations. Let $Z \in \mathbb{R}^{b\times d\times h\times w}$ be the feature map before the CAW layer, where b, d, h, and w denote the batch size, dimension, height, and width. The CAW layer is composed of two operations: (1) a whitening

transformation to separate different concepts in the latent space; and (2) an orthogonal transformation to align the axes of latent space with pre-defined concepts. Next, we will introduce each operation in detail.

Whitening transformation. First, we flatten the feature map $Z \in \mathbb{R}^{b\times d\times h\times w}$ into shape $d \times n$, where $n = b \times h \times w$. Then, a whitening transformation ψ is adopted to decorrelate and standardize the feature Z by:

$$\psi(Z) = W(Z - \mu \mathbf{1}_{1\times n}), \tag{1}$$

where μ is the mean of n samples; $W \in \mathbb{R}^{d\times d}$ is the whitening matrix which can be calculated by ZCA algorithm [11]. After whitening, each dimension of the feature becomes mutually independent.

Orthogonal transformation. In this step, we align each separated dimension to a specific concept. This is achieved by leveraging an orthogonal matrix $Q \in \mathbb{R}^{d\times d}$, in which the column q_k is defined as the feature of concept c_k. For the calculation of the matrix Q, we particularly propose a weakly-supervised concept alignment, which will be elaborated in the next section. As the whitening matrix W is rotation-free, $Q^T W$ is also a valid whitening matrix. Thus after CAW, we can obtain the interpretable feature $Z' = Q^T\psi(Z)$ and reshape it into its original size $b \times d \times h \times w$ for subsequent computation.

Finally, the feature map Z' is fed into the rest of the network to predict the disease label by the following objective:

$$\min \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{ce}(g(Q^T\psi(f(x_i))), y_i), \tag{2}$$

where f and g are layers before and after the CAW layer respectively. ψ is a whitening transformation. Q is the orthogonal matrix. $\mathcal{L}_{ce}$ is the cross-entropy loss for skin disease classification.

2.2 Concept Alignment Branch

The concept alignment branch aims to estimate an orthogonal matrix Q for the disease diagnosis branch by leveraging the concept dataset X_c. Specifically, the concept dataset $X_c = \{X_{c_k}\}_{k=1}^{K}$ consists of K subsets of images, where each X_{c_k} represents the images with concept c_k. We develop a weakly-supervised concept mask generator to identify concept-attentive features, which serve as guidance for refining the orthogonal matrix Q.

Weakly-supervised concept mask generator. First, we pre-train a concept classification network on the concept dataset X_c supervised by concept labels. We generate concept activation maps from the pre-trained concept classification network in a weakly-supervised manner. Formally, given a concept feature map $\mathbf{Z}^{\text{con}} \in \mathbb{R}^{d\times h\times w}$, the weights of the classifier $\mathbf{W} \in \mathbb{R}^{d\times K}$ can be regarded as the prototypes of K concepts. With regard to the predicted concept class c_k, we select the corresponding prototype $\mathbf{W}^{c_k} \in \mathbb{R}^d$ and measure its similarity with each pixel of the feature map $\mathbf{Z}^{\text{con}}$. The activation value of $\mathbf{M}_{c_k}(i, j)$ at spatial

location (i, j) is calculated by summing the multiplication of $\mathbf{W}^{c_k}$ and $\mathbf{Z}^{\text{con}}(i, j)$ across the channel dimension:

$$\mathbf{M}_{c_k}(i, j) = \sum_d \mathbf{W}_d^{c_k} \cdot \mathbf{Z}_d^{\text{con}}(i, j). \tag{3}$$

The activation map $\mathbf{M}_{c_k} \in \mathbb{R}^{h \times w}$ is normalized to the interval $[0, 1]$. We further binarize the concept map by a pre-defined threshold γ to generate the concept mask, which is used to filter the most discriminative concept features:

$$\tilde{\mathbf{M}}_{c_k}(i, j) = \begin{cases} 1, & \text{if } \mathbf{M}_{c_k}(i, j) > \gamma \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

The concept mask $\tilde{\mathbf{M}}_{c_k} \in \{0, 1\}^{h \times w}$ highlights one specific concept of interest c_k in the image, and it is subsequently used to guide the optimization of Q.

Optimization of Q. To align the k-th feature dimension with concept c_k, we need to find an orthogonal matrix $Q \in \mathbb{R}^{d \times d}$ whose column q_k corresponds to the k-th axis, by optimizing the concept alignment objective:

$$\begin{gathered} \max_{q_1, q_2, \ldots, q_K} \sum_{k=1}^{K} \frac{1}{|X_{c_k}|} \sum_{x_{c_k} \in X_{c_k}} q_k^T \text{AvgPool}(\tilde{\mathbf{M}}_{c_k} \odot \psi(f(x_{c_k}))), \\ s.t\ Q^T Q = I_d \end{gathered} \tag{5}$$

where AvgPool denotes average pooling over the spatial dimension, resulting in a concept-attentive feature vector with shape $d \times 1$. This optimization problem is a linear programming problem with quadratic constraints (LPQC), which is generally NP-hard. Since directly solving the optimal solution is intractable, we optimize it by gradient methods on the Stiefel manifold [24]. At each step t, the orthogonal matrix Q is updated by Cayley transform:

$$Q^{(t+1)} = (I + \frac{\eta}{2} A)^{-1} (I - \frac{\eta}{2} A) Q^{(t)}, \tag{6}$$

where η is the learning rate, $A = G(Q^{(t)})^T - Q^{(t)} G^T$ is a skew-symmetric matrix and G is the gradient of the loss function.

3 Experiments

3.1 Datasets and Implementation Details

Datasets. Derm7pt [12] consists of 1,011 dermoscopic images, annotated with disease and clinical concept labels. Following [19, 2], we filter the dataset to obtain a subset of 827 images. The disease categories include nevus and melanoma, and the concept categories cover 12 elements from the 7-point checklist [1]. SkinCon [4] includes 3,230 images from the Fitzpatrick 17k skin disease dataset [9] that are densely annotated with 48 clinical concepts. The disease categories comprise malignant, benign, and non-neoplastic. We select 22 concepts that have at

Table 1. Disease diagnosis results of concept-based state-of-the-art methods. We report the results as mean_{std} of three random runs.

Method	Derm7pt			SkinCon		
	AUC	ACC	F1	AUC	ACC	F1
ResNet [10]	$89.48_{0.46}$	$84.48_{0.47}$	$80.60_{0.69}$	$80.85_{0.71}$	$78.85_{0.57}$	$77.80_{0.64}$
Sarkar et al. [23]	$76.22_{2.06}$	$73.89_{1.47}$	$66.81_{1.23}$	$68.21_{1.44}$	$71.14_{1.21}$	$71.32_{1.38}$
PCBM [25]	$72.96_{2.19}$	$76.98_{1.39}$	$71.04_{1.15}$	$68.94_{1.59}$	$71.04_{1.13}$	$70.47_{0.75}$
PCBM-h [25]	$83.27_{1.14}$	$79.89_{0.89}$	$74.48_{1.37}$	$69.53_{1.67}$	$72.28_{1.39}$	$72.28_{1.29}$
CBE [19]	$76.60_{0.35}$	$83.75_{0.26}$	$78.13_{0.44}$	$72.75_{1.15}$	$73.75_{1.10}$	$73.56_{1.31}$
MICA w/ bot [2]	$84.11_{1.10}$	$82.20_{1.31}$	$78.08_{1.22}$	$75.89_{1.11}$	$74.29_{1.09}$	$74.74_{1.21}$
MICA w/o bot [2]	$85.59_{1.11}$	$83.94_{0.99}$	$79.38_{1.34}$	$75.92_{1.13}$	$75.63_{1.07}$	$75.43_{1.24}$
CW [3]	$86.50_{0.40}$	$83.85_{0.48}$	$80.00_{0.75}$	$79.49_{0.60}$	$78.28_{0.57}$	$77.30_{0.67}$
CAW (ours)	$\mathbf{88.60}_{0.10}$	$\mathbf{84.79}_{0.79}$	$\mathbf{81.34}_{0.85}$	$\mathbf{80.47}_{0.24}$	$\mathbf{79.00}_{0.19}$	$\mathbf{77.76}_{0.57}$

least 50 images representing the concept. The dataset is partitioned into 70%, 15%, and 15% for training, validation, and testing, respectively. The specific concepts used in the two datasets are enumerated in the Supplemental Material.

Implementation details. For the Derm7pt [12] and SkinCon [4] datasets, we employ the ImageNet-pretrained ResNet-18 and ResNet-50 models [10] as the backbones, respectively. We replace the BN with CAW layer in the 8/16-th layer for ResNet-18/50. All images are resized to 224×224. Data augmentation includes random horizontal flip, random cropping, and rotation. The network is trained for 100 epochs with a batch size of 64 and a learning rate of 2e-3. We adopt AUC, ACC, and F1 score as the evaluation metrics.

3.2 Experimental Results

Skin disease diagnosis. We compare the performance of skin disease diagnosis of our proposed CAW with other state-of-the-art methods, as shown in Table 1. To establish an upper bound, we first train a standard black-box ResNet model without the interpretability of concepts. In comparison, the group of CBM-based models [23, 25, 19, 2] demonstrates a significant performance decrease for skin disease diagnosis, despite enhancing interpretability. Notably, the vanilla CW model [3] shows superior diagnosis performance, surpassing all the CBM-based methods across all the evaluation metrics. With concept attention, our proposed CAW significantly further improves the skin disease diagnosis performance on both datasets. CAW approaches the performance of the black-box model in terms of AUC, and even surpasses it on ACC and F1.

Concept detection. We conduct concept detection to measure the interpretability of our CAW quantitatively. Following [3], we calculate the one-vs-all test AUC score of classifying the target concept in the latent space. We compare our CAW with the concept vectors learned by TCAV [15] from black-box models, the filters in standard CNNs [26], and CW [3]. As illustrated in Fig. 2, our CAW outperforms all the other methods on the average AUC score, reaching 77.4%

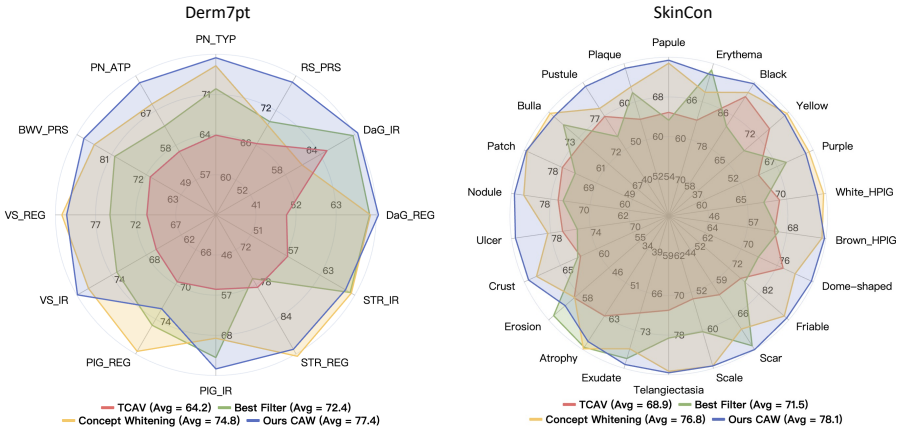

Fig. 2. Comparison of concept detection on Derm7pt and SkinCon.

on Derm7pt and 78.1% on SkinCon. Moreover, our CAW demonstrates superior or comparable detection performance on most concepts.

3.3 Ablation Study and Discussion

Effect of concept attention mask. We conduct an ablation study to investigate the effect of our generated concept mask. As shown in the first row of Table 2, the baseline approach, which merely uses the raw image as the concept map, yields the poorest performance, as it fails to capture precise concept features. The random and center-gaussian maps appear to enhance the performance, which can be attributed to their roles as a source of Cutout augmentation [5]. To take a step further, we trained a skin lesion segmentation model [22] to generate lesion masks, effectively eliminating redundant regions in the images and emphasizing the entire lesion area, leading to improved performance. However, there still exists a difference between the lesion area and the concept regions. Finally, our generated concept masks by thresholding on concept maps accurately localize concept regions, resulting in the best performance in terms of both disease diagnosis and concept detection.

Table 2. Ablation study on concept mask.

Method	Disease Diag.	Concept Det.
raw image	86.96	74.79
gaussian map	87.16	74.92
random map	87.78	75.93
lesion mask	87.94	75.55
concept map	88.13	76.23
concept mask	**88.60**	**77.40**

Table 3. Analysis on threshold.

γ	Disease Diag.	Concept Det.
0	86.95	74.79
0.2	87.68	77.12
0.5	<u>88.60</u>	**77.40**
0.6	88.44	<u>77.21</u>
0.8	**88.63**	76.52
1.0	87.92	76.03

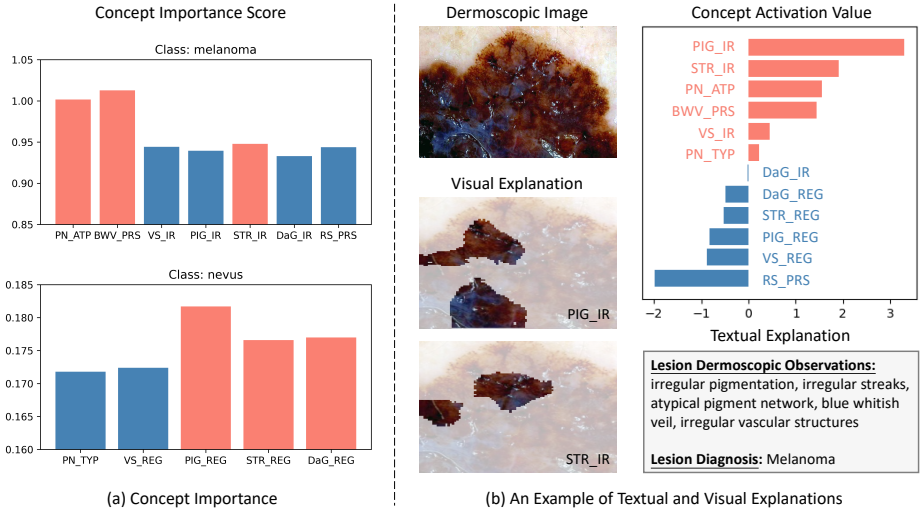

Fig. 3. (a) Analysis on concept importance. (b) An example of explanations.

Analysis on threshold. We also investigate the impact of different binarization threshold values for concept mask generation. The results in Table 3 demonstrate that the performance consistently maintains a high level within the intermediate range of 0.5 to 0.8, indicating the robustness of our model to the choice of threshold. It is believed that a very small threshold would lead to the presence of concept-irrelevant regions, while a large threshold results in the loss of potential key concept information.

Concept importance. We measure the contribution of concepts to the disease diagnosis. The importance of the k-th concept CI_k is defined as the ratio between the switched loss and the original loss [8], which is given by $CI_k = loss^k_{switch}/loss_{orig}$. Here, $loss_{orig}$ denotes the original loss produced by the network without any permutation. To calculate the switched loss $loss^k_{switch}$, we randomly permute the k-th value of the concept feature along the batch dimension, *i.e.*, replacing the k-th concept value of the current sample with another one from the batch. This indicates that the switched loss is expected to be large if the k-th concept is important for the current sample. We show the concept importance scores on Derm7pt. The results depicted in Fig. 3 (a) indicate the top three important concepts for melanoma and nevus, respectively, which confirm the consistency with the findings of dermatologists [18].

Example of explanation. We present an example that demonstrates how our CAW can offer comprehensible explanations during the disease diagnosis process, as illustrated in Fig. 3 (b). For an input image, the activation value of its representation can be interpreted as the probability associated with a certain concept. Based on the activation values, a textual explanation can be derived to

describe the image, and visual explanations can be generated simultaneously to emphasize the specific concept regions.

4 Conclusion

In this work, we propose an intrinsic interpretable XAI model based on concept, *i.e.*, Concept-Attention Whitening (CAW), for skin lesion diagnosis. CAW consists of a disease diagnosis branch and a concept alignment branch. We specifically incorporate a weakly-supervised concept mask generator to filter the most relevant local regions, benefiting precise optimization of the orthogonal matrix in the CAW layer. Experiments on two skin lesion diagnosis datasets demonstrated the interpretability and superior diagnostic performance of CAW. In future work, we will consider the correlation of concepts by softening the orthogonality constraints in CAW, which is expected to promote the discovery of new concepts.

Acknowledgments. This work was supported by the HKUST (Project No. FS111) and Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083).

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

References

1. Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., Delfino, M.: Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Archives of dermatology **134**(12), 1563–1570 (1998)
2. Bie, Y., Luo, L., Chen, H.: Mica: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. arXiv preprint arXiv:2401.08527 (2024)
3. Chen, Z., Bei, Y., Rudin, C.: Concept whitening for interpretable image recognition. Nature Machine Intelligence **2**(12), 772–782 (2020)
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. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
6. 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)
7. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)

8. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**(177), 1–81 (2019)
9. 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)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
11. Huang, L., Zhou, Y., Zhu, F., Liu, L., Shao, L.: Iterative normalization: Beyond standardization towards efficient whitening. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4874–4883 (2019)
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 (mar 2019). https://doi.org/10.1109/JBHI.2018.2824327
13. Kazhdan, D., Dimanov, B., Jamnik, M., Liò, P., Weller, A.: Now you see me (cme): concept-based model extraction. arXiv preprint arXiv:2010.13233 (2020)
14. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
16. Kim, E., Jung, D., Park, S., Kim, S., Yoon, S.: Probabilistic concept bottleneck models. arXiv preprint arXiv:2306.01574 (2023)
17. 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)
18. Menzies, S., Ingvar, C., McCarthy, W.: A sensitivity and specificity analysis of the surface microscopy features of invasive melanoma. Melanoma research **6**(1), 55–62 (1996)
19. 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. 3798–3807 (2023)
20. Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., Baralis, E.: Concept-based explainable artificial intelligence: A survey. arXiv preprint arXiv:2312.12936 (2023)
21. Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., Scotton, P.: Attention-based interpretability with concept transformers. In: International Conference on Learning Representations (2021)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
23. Sarkar, A., Vijaykeerthy, D., Sarkar, A., Balasubramanian, V.N.: A framework for learning ante-hoc explainable models via concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10286–10295 (2022)

24. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Mathematical Programming **142**(1-2), 397–434 (2013)
25. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2022)
26. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6856