Towards Trustable Skin Cancer Diagnosis via Rewriting Model's Decision

Siyuan Yan[1,2] Zhen Yu[1,2] Xuelin Zhang[1,2] Dwarikanath Mahapatra[4]
Shekhar S. Chandra[3] Monika Janda[3] Peter Soyer[3] Zongyuan Ge[1,2]
[1] Monash University [2] Monash Medical AI Group [3] The University of Queensland
[4] Inception Institute of AI, Abu Dhabi, UAE

Abstract

Deep neural networks have demonstrated promising performance on image recognition tasks. However, they may heavily rely on confounding factors, using irrelevant artifacts or bias within the dataset as the cue to improve performance. When a model performs decision-making based on these spurious correlations, it can become untrustable and lead to catastrophic outcomes when deployed in the real-world scene. In this paper, we explore and try to solve this problem in the context of skin cancer diagnosis. We introduce a human-in-the-loop framework in the model training process such that users can observe and correct the model's decision logic when confounding behaviors happen. Specifically, our method can automatically discover confounding factors by analyzing the co-occurrence behavior of the samples. It is capable of learning confounding concepts using easily obtained concept exemplars. By mapping the black-box model's feature representation onto an explainable concept space, human users can interpret the concept and intervene via first order-logic instruction. We systematically evaluate our method on our newly crafted, well-controlled skin lesion dataset and several public skin lesion datasets. Experiments show that our method can effectively detect and remove confounding factors from datasets without any prior knowledge about the category distribution and does not require fully annotated concept labels. We also show that our method enables the model to focus on clinical-related concepts, improving the model's performance and trustworthiness during model inference.

1. Introduction

Deep neural networks have achieved excellent performance on many visual recognition tasks [11, 14, 35]. Meanwhile, there are growing concerns about the trustworthiness of the model's black-box decision-making process. In medical imaging applications, one of the major issues is deep learning models' confounding behaviors using irrelevant artifacts (*i.e.* rulers, dark corners) or bias (*i.e.* image backgrounds, skin tone) as the cue to make the final predictions. These spurious correlations in the training distribution can make models fragile when novel testing samples are presented. Therefore, transparency of decision-making and human-guided bias correction will significantly increase the reliability and trustworthiness of model deployments in a life-critical application scenario like cancer diagnosis.

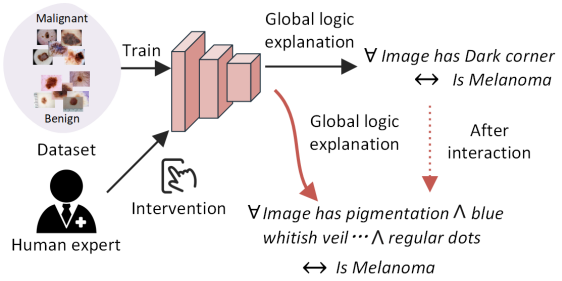


Figure 1. Our method allows people to correct the model's confounding behavior within the skin lesion training set via rewriting the model's logic.

For instance, due to the nature of dermatoscopic images, the skin cancer diagnosis often involves confounding factors [4, 22, 24, 40], *i.e.*, dark corners, rulers, and air pockets. Bissoto *et al.* [40] shows that deep learning models trained on common skin lesion datasets can be biased. Surprisingly, they found the model can reach an AUC of 73%, even though lesions in images are totally occluded. To better understand the issue of confounding behaviors, we illustrate a motivating example. Fig. 1 shows a representative confounding factor "dark corners" in the *ISIC2019-2020* [36, 38] dataset, where the presence of dark corners is much higher in the melanoma images than in benign images. A model trained on this dataset tends to predict a lesion as melanoma when dark corners appear. This model is undoubtedly untrustable and is catastrophic for deployment in real clinical practice. To deal with this nuisance and improve the model's trustworthiness, an ideal method would be: i) the model itself is explainable so that humans can understand the prediction logic behind it. and ii) the model has an intervening mechanism that allows humans to correct the model once confounding behavior happens. Through this

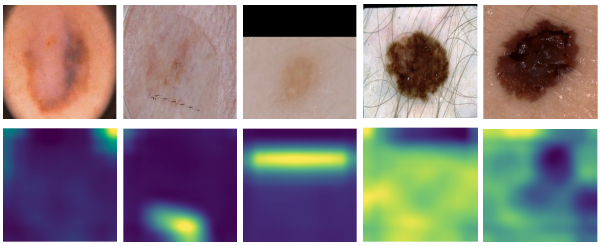
way, the model can avoid dataset biasing issues and rely on expected clinical rules when needed to diagnose.

In this work, our overall goal is to improve model transparency as well as develop a mechanism that human-user can intervene the model training process when confounding factors are observed. The major difference between our proposed solution and existing works [19,28,29,33] are: (1) For **model visualization**, we focus on generating human-understandable textual concepts rather than pixel-wise attribution explanations like Class Activation Maps [42]. (2) For **concept learning**, we do not hold any prior knowledge about the confounding concepts within the dataset. This makes our problem set-up more realistic and practical than the existing concept-based explanation or interaction [19, 33] framework where fully-supervised concept labels are required. (3) For **method generalization**, our method can be applied on top of any deep models.

Therefore, we propose a method that is capable of learning confounding concepts with easily obtained concept examplars. This is realized via clustering model's co-occurrence behavior based on spectral relevance analysis [20] and concept learning based on concept activation vector (CAVs) [18]. Also, a concept space learned by CAVs is more explainable than the feature-based counterparts.

To the end, we propose a human-in-the-loop framework that can effectively discover and remove the confounding behaviors of the classifier by transforming its feature representation into explainable concept scores. Then, humans are allowed to intervene based on first-order logic. Through human interaction (see Fig. 1) on the model's concept space, people can directly provide feedback to the model training (feedback turned into gradients) and remove unwanted bias and confounding behaviors. Moreover, we notice that no suitable dermatology datasets are available for confounding behavior analysis. To increase the methods' reproducibility and data accessibility in this field, we curated a well-controlled dataset called *ConfDerm* containing 3576 images based on well-known *ISIC2019* and *ISIC2020* datasets. Within the training set, all images in one of the classes are confounded by one of five confounding factors, including dark corners, borders, rulers, air pockets, and hairs. Still, in the testing set, all images are random. Some confounding factors within the dataset are illustrated in Fig. 2.

We summarize our main contributions as: (1) We crafted a novel dataset called ConfDerm, which is the first skin lesion dataset used for systematically evaluating the model's trustworthiness under different confounding behaviors within the training set. (2) Our new spectral relevance analysis algorithm on the popular skin cancer dataset *ISIC2019-2020* has revealed insights that artifacts such as dark corners, rulers, and hairs can significantly confound modern deep neural networks. (3) We developed a human-in-the-loop framework using concept mapping and logic layer rewriting to make the model self-explainable and enable users effectively remove the model's confounding behaviors. (4) Experimental results on different datasets demonstrate the superiority of our method on performance improvement, artifact removal, and skin tone debiasing.



(a) (b) (c) (d) (e)

Figure 2. Observed confounding concepts in *ISIC2019-2020* datasets, the top row shows sample images, and the bottom row is the corresponding heatmap from GradCAM: (a) dark corners. (b) rulers. (c) dark borders. (e) dense hairs. (f) air pockets.

2. Related Work

2.1. Trustable Skin Diseases Diagnosis

In dermatology, clinicians usually diagnose skin cancer by assessing skin lesions based on criteria such as the ABCD rule [2] and the seven-point checklist [1]. As visual examination for skin cancer diagnosis is subjective and relies on the experience of dermatologists, deep learning-based diagnostic algorithms have been gaining attraction. Recently, Deep Learning methods *i.e.*, DeepDerm [11], have reached dermatologist-level diagnostic accuracy. However, the black-box nature of deep learning methods in the decision-making process has been a big hurdle to making clinicians trust them. Several researchers try to make skin cancer diagnosis trustable based on either attribution-based explanation methods (*i.e.*, Integrated Gradients [34], GradCAM [31], LIME [23]) or concept-based explanation methods (*i.e.*, Concept bottleneck Model [19], Concept Activation Vector [18], and Posthoc Concept Model [41]). Our work follows the concept-based explanations but is different from them in the following aspects: (1) Existing concept-based models for skin disease diagnosis only consider clinically related concepts and do not have the ability to find spurious correlations within the training data, making them can not conduct fine-grained error analysis. (2) we further studied using interactive learning to remove confounding behavior, which is an important but rarely explored point in skin cancer diagnosis and is also an essential step in building trustworthiness between clinicians and models.

2.2. Concept-based Explanations

Concept-based explanations identify high-level semantic concepts that apply to the entire dataset that is not limited by

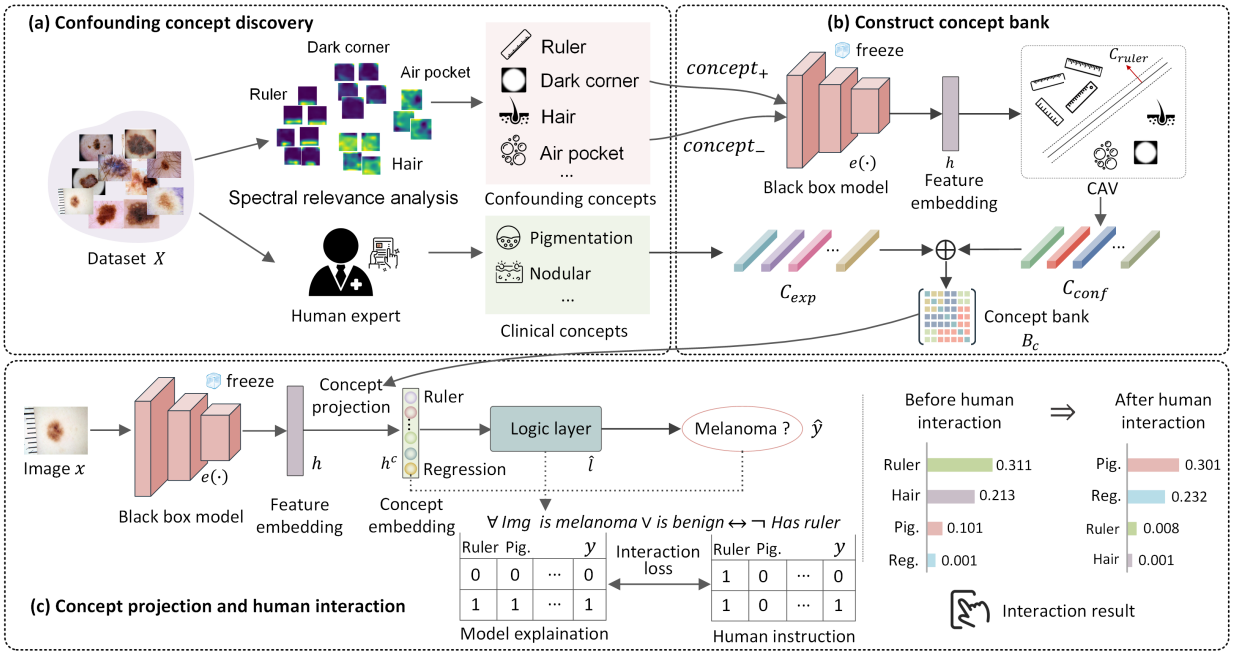


Figure 3. Illustration of our human-in-the-loop pipeline. (a) Applying the improved spectral relevance analysis algorithm to discover the confounding factors within the dataset (see subsection 3.1). (b) Learning confounding concept and clinical concept vectors (see subsection 3.2. (c) Projecting feature representation of a model onto concept subspace and then removing the model's confounding behaviors via human interaction (see subsection 3.3). Pig. denotes pigmentation, and Reg. denotes regression structure.

the feature space of deep neural networks. It can be roughly divided into two categories, concept bottleneck network (CBM) like methods [3, 19, 27], and Concept Activation Vectors(CAVs) like methods [18, 21, 41]. The CBM, as an interpretability-by-design model, first predicts the concepts, then predicts the target labels using the concepts by restricting the model's architecture. We argue that this method is not suitable for skin cancer diagnosis since CBM assumes the relevance between concepts and tasks are known in advance, but it is not the case when we want to solve new tasks. Also, the per-image level concept annotation is very expensive, especially for skin cancer datasets requiring expert annotation. Different from CBM, the CAVs train linear classifiers to the neural network's feature representation to verify whether the representation can separate the concept examples defined by people. Then, the coefficients of the linear classifier are CAVs. With the CAVs, we can get global explanations for a model's overall behaviors. As the user-defined concept examples are relatively easily obtained for different tasks (only need 20-70 images for each concept as the exemplar), we make use of CAVs in our work.

2.3. Explanatory Interactive Learning

Explanatory Interactive Learning (XIL) is a combination setting of Explainable AI (XAI) and active learning. It incorporates the human user in the training loop by querying the model's explanations and then giving feedback to correct the model's explanations. Some works [9, 15] tend to force the model to attend to important regions in the input images via attention learning, but attention is often different from explanation [9, 16]. Another line is to constrain the model's explanation by penalizing the model's gradient outside the correct regions to align human knowledge. [26, 30] achieve it by providing the model's human-annotated masks during training. However, it is too expensive to pixel-wise annotate every image in skin lesion datasets. [33] tends to compile a lot of concept labels for simulated datasets and changes the model's behavior via intervention on its concept representation of the model. Unfortunately, this method can not be applied to skin cancer diagnosis as it requires us to know and annotate all possible concepts in our task; this is not possible for real-world cases.

3. Method

To formalize our problem, consider a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ where x_i is an input image, y_i is the corresponding ground truth, and N is the size of the training dataset, and a CNN classification model $f = l \circ e$ where the feature extractor e maps x_i into latent embedding h_i and then the classifier l maps the h_i into final prediction $y_i = l(h_i)$. Our objective is to identify the confounding factors $C_{conf} = \{(c_j)\}_{j=1}^{M}$ within the training set $\mathcal{D}$ that can confound f, and then remove them via XIL on the ex-

plainable version of f.

The overall pipeline of our method is summarized in Fig. 3. First, we apply spectral clustering [39] on the GradCAM [31] computed with f to discover different confounding concepts within the dataset D (see Fig. 3 (a) and detail in section 3.1). Then, we construct a concept bank B_c, which is composed of confounding concepts C_{conf} and clinical-related expert concepts C_{exp}. We learn concept activation vectors (CAVs) from different clusters to obtain C_{conf}, and learn CAVs from a small probe dataset with expert-level concept annotations to obtain C_{exp} (see Fig. 3 (b) and section 3.2). Next, we project the feature representation h from extractor e onto a concept subspace h^c spanned by B_c. We replace the l with an explainable logic layer $\hat{l}$ to model the relationship between concepts and the final prediction $\hat{y} = \hat{l}(h)$ based on the concept representation h^c. Finally, we rewrite the model's decision during training by applying interaction loss on the input gradient of $\hat{l}$. The details are elaborated in Fig. 3 (c) and section 3.3.

3.1. Global Confounding Concepts Discover

In this section, we introduce our global confounding concept discovery algorithm (GCCD), which is used to semi-automatically discover the model's common confounding behaviors within the training set. Our method is based on spectral relevance analysis (SpRAy) [20], originally used to analyze the co-occurring behavior with a dataset. As illustrated in Algorithm 1, given a training set X, class labels C, and a model f trained on the X, we want to visualize and obtain concept clusters in a 2D embedding E. First, we randomly sample a subset $\hat{X}$ from X. Then, we calculate the GradCAM heatmaps of images from $\hat{X}$ for each class. Instead of only performing prepossessing and normalization on heatmaps like SpRAy, we additionally apply discrete Fourier Transformation [5] to distinguish different models' behaviors better to obtain high-quality concept clusters for similar but different concepts. Then, we concatenate each heatmap with its corresponding image to provide additional appearance information, which is different from SpRAy, which highly relies on the location information of the heatmaps. To accelerate the clustering process, we downscale the concatenated images five times, as suggested in [20]. Similar to [20, 29], we choose spectral clustering [39] to calculate a spectral embedding based on an adjacency matrix calculated by the K-nearest neighborhood algorithm (KNN). To analyze the clusters, we perform non-linear dimension reduction via t-SNE [37]. Finally, we manually annotate the concept label for each cluster and filter out those non-representative clusters.

3.2. Construction of Concept Bank

To build our concept bank C_{conf}, we make use of Concept Activation Vectors (CAVs) [18]. For each concept, given a series of positive concept examples $P^c = \{P_i^c\}_{i=1}^{T_P}$ and negative concept examples $N^c = \{N_i^c\}_{i=1}^{T_N}$. We train a linear classifier to separate the CNN features of those examples that contain the concept $e(P^c)$ or not $e(N^c)$. Then, the CAVs are defined as a normal w^c to a hyperplane that separating $e(P^c)$ from $e(N^c)$ in the embedding space h, such that satisfying $(w^c)^T h + \phi^c > 0$ for all $h \in e(P^c)$ and $(w^c)^T h + \phi^c < 0$ for all $h \in e(N^c)$ where w^c and ϕ^c is the weight and bias of the linear classifier.

For our case, we collect confounding concepts C_{conf} from previously generated concept clusters E with a human confirmation. To collect clinical-related concepts C_{exp}, we collect concepts from existing expert-level annotated probe datasets (*i.e.*12 concepts from *Derm7pt* [17] for dermatoscopic images or 22 concepts from *skincon* [8] for clinical images). Finally, we get the concept bank $B_c = C_{conf} \oplus C_{exp}$; more details will be described in the Experimental part.

Algorithm 1: Global Confounding Concept Discovery (GCCD).

Input: Training set $X = \{x_i\}_{i=1}^{N}$
Class labels $C = \{c_j\}_{i=1}^{M}$
Model f
Output: t-SNE embedding $E = \{e_k\}_{k=1}^{M}$
Randomly sample a subset of X as $\hat{X}$;
for $C_j \in C$ **do**
  $M_{cj} = \{\}$;
  for $\hat{x_i} \in \hat{X}$ **do**
    $M_j = GradCAM(f, \hat{x}_i, c_j)$;
    $M_j = FourierTransform(M_i)$;
    $\hat{x_i} = Preprocessing(\hat{x_i})$;
    $M_j = M_j \oplus \hat{x_i}$;
    $M_j = Downscale(M_j)$;
    $M_{cj} \leftarrow M_j$;
  end
  Obtain adjacency matrix $H_j = KNN(M_{cj})$;
  Spectral clustering on H_j, get the sepctral embedding ϕ_j;
  Visualize 2D embedding $e_k = tSNE(\frac{1}{H_j + \epsilon})$;
end
return E

3.3. Model Logic Rewriting

Turn the black box into explainable models: A black-box model f (could be any backbone *i.e.*, ResNet, Inception, or Vision transformer) can be simplified as a composition of a feature extractor e, and classification decision layers l. Our aim is to make it logically interpretable to increase its trustworthiness. First, we project hidden embedding features h of $e(X)$ onto a low dimensional concept embedding h^c. Each dimension of h^c corresponds to a concept in B_c. Specifically, we define $h^c = \frac{\langle h, B_c \rangle}{\|B_c\|^2} B_c$ where h^c is concept scores, B_c is the concept bank containing CAVs for all concepts. We replace the l with an explainable layer $\hat{l}$, so that $\hat{l}(h^c) = \hat{y}$ where its prediction $\hat{y}$ is based

on the explainable concept embedding h^c. The linear layer, decision tree, or logic layer can be used as the explainable layer $\hat{l}$. In the medical community, dermatologists diagnose a lesion based on combining prediction results for different clinical concepts (*i.e.*ABCD rule [2] and seven-point checklist [1]) to get a diagnosis score. A lesion is diagnosed as melanoma if its diagnosis score is larger than a threshold. To mimic the skin cancer diagnosis process and produce human understandable interpretability, we choose the recently proposed entropy-based logical layer [3] as it can binarize the concepts and model their relationship based on importance weights to perform final decision. More specifically, it produces first-order logic-based explanations by performing an attention operation on the model's concept representation. However, [9, 16] show that attention is often not the explanation, which causes interaction on it is not effective in changing the model's behavior (see examples and module details in supplementary material).

Interaction and model re-writing: To generate faithful explanations of the logic layer $\hat{l}$, we employ input gradient explanation to highlight the important concepts in the concept representation h^c. Given $h^c \in b \times d$ and $\hat{l} \in b \times d \rightarrow b \times p$ where b is the number of samples, d is the number of concepts, and p is the number of classes, the input gradient is defined as $\nabla_{h^c}\hat{l}(h^c)$, which is a vector norm to the model's decision boundary at h_i^c, and can be regarded as a first-order description of model's behavior around h_i^c. The users are able to intervene in the model's behavior by providing an annotation matrix $H \in b \times d$ to model $\hat{l}$, where H represents the first-order logic used to inform the model which concepts are irrelevant to its prediction. We achieve this by employing the right for the right reason (RRR) loss [28]. We replace the $L2$ regularization term with the elastic-net regularizer to improve the model's sparsity, and therefore users can more easily intervene with its explanation. Our final loss is defined as:

$$\mathcal{L}(h^c, y, H) = \underbrace{\mathcal{L}_{CE}(y, \hat{y})}_{\text{Cross Entropy Loss}} + \underbrace{\lambda_1(\alpha\|w\|_1 + (1-\alpha)\|w\|_2^2)]}_{\text{Elastic-Net Regular}} + \underbrace{\lambda_2 \sum_{b=1}^{B}\sum_{d=1}^{D}(H_{bd}\frac{\partial}{\partial h_{bd}^c}\sum_{c=1}^{N^c}(\hat{y}_{bc}))^2}_{\text{Right Reasons Loss}} \quad (1)$$

where the right reason loss penalizes the input gradient from being large in the positions annotated by a human user (see the graph at the bottom of Fig. 3). The λ_1 controls the sparsity of $\hat{l}$, and λ_2 controls how much the human user instruction H is considered by the model.

4. ConfDerm Dataset

We found it difficult to evaluate the model's confounding behavior on existing dermatology datasets due to two reasons: (1) removing confounding behaviors within a dataset does not always improve performance, especially when the distribution in the testing set is similar to the training set (*i.e.*, both have dark corners). (2) It is also hard to quantify all confounding factors in a real skin lesion dataset. To alleviate these problems, we collect a well-controlled dataset *ConfDerm* with 3576 real dermoscopic images with melanoma and benign classes from *ISIC2019* [36], and *ISIC2020* [38]. Our dataset consists of five sub-datasets to capture different confounding factors for a particular class, including *melanoma (dark corners)*, *benign (dark borders)*, *benign (rulers)*, *benign (hairs)*, and *melanoma (air pockets)*. For example, in the *benign (dark borders)* dataset, all benign images contain dark borders in the training set but no dark borders in the testing set. Still, all benign images in the training and testing set are random. The concept distribution of all five datasets is described in Fig. 4, and we omit the class distribution of each sub-dataset as classes in all datasets are balanced.

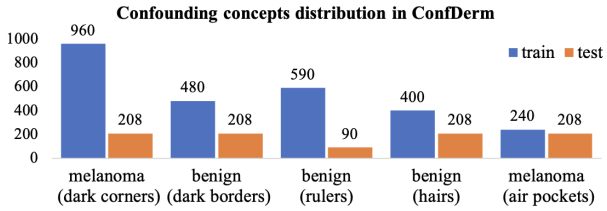


Figure 4. Confounding concepts distribution in our proposed *ConfDerm* dataset.

5. Experiments

5.1. Experimental setup

Our task is to classify a skin lesion into benign or melanoma. To evaluate the trustworthiness of the above task of using conventional CNN vs. our proposed method, we perform three main experiments on five public training sets and one novel dataset we crafted: **(1) confounding concepts discovery**: we use our GCCD (see Algorithm.1) to discover the confounding factors on three popular skin lesion datasets and one synthetic dataset. **(2) rewriting model's decision:** we perform our human interaction method on our novel dataset *ConfDerm* to evaluate its effectiveness in handling five challenging confounding problems of dermoscopic images. **(3) Debiasing the Effect of Skin Tone:** we also study the effect of skin tone using our interaction method. We evaluate model robustness across datasets with different skin tone scales by removing the skin tone concept.

Training setting: Our method is general and can be applied to any backbone, such as VGG [32], ResNet [14], Inception [35], and ViT [10]. For subsection 5.2 and 5.3, we use ResNet50 as the backbone model, and for subsection 5.4, following the setting in [7, 11], we use Incep-

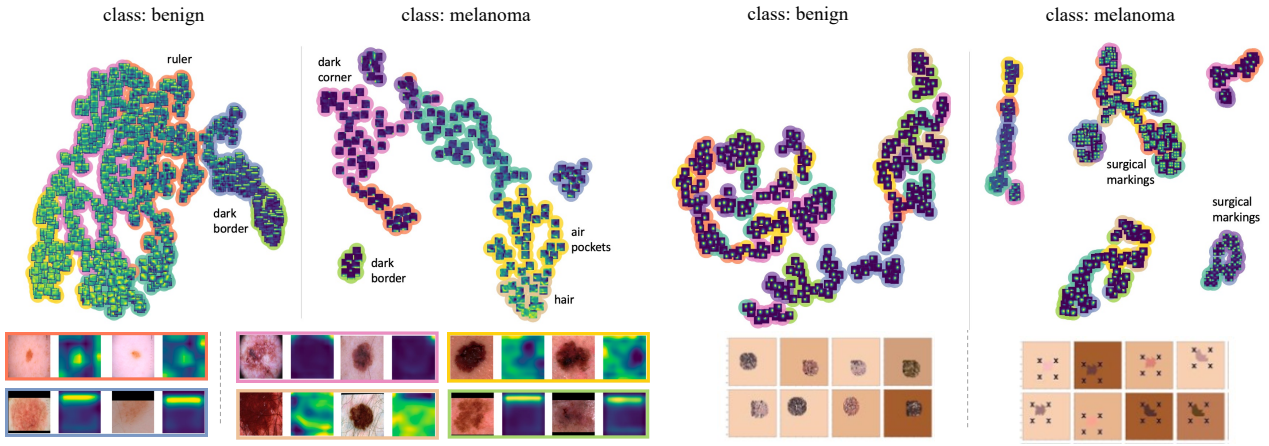


Figure 5. Global analysis of the models' behavior within datasets using GCCD. The left graph is the tSNE of spectral clustering using GradCAMs of a ResNet50 within *ISIC2019* and *ISIC2020*. The right one is the tSNE of spectral clustering using GradCAMs of a VGG16 within the *SynthDerm* dataset. More detailed visualization is included in the supplementary material.

tion V3 with DeepDerm weights as our backbone and fine-tune it on *Fitzpatrick 17K* dataset. For evaluation metrics, we use accuracy for experiments in subsection 5.3 as the dataset *ConfDerm* has balanced classes, and we use ROC-AUC for all other skin cancer diagnosis experiments. For hyper-parameters, prepossessing and more detail of each used dataset is in the supplementary material.

Human interaction: Most existing XIL settings [25, 29] either use per-sample level human-annotated masks or concept labels to simulate the human user to have full knowledge about the task. However, it is often not realistic and general for skin cancer diagnosis as most skin lesion datasets do not have these annotations, and these annotations are also very expensive. Different from these local interaction methods, we choose to use global interaction similar to [33] on confounding factors. For example, we provide a rule table defined by human-user in the column of the concept "rulers" is all 1s, and other columns are all 0s, which means "never focus on rulers" (see Fig. 3 (c)). This is an efficient method to perform interaction as we only need one rule table to change the model's behavior for the entire dataset. More details about interaction are described in the following subsections.

5.2. Confounding Concept Discovery

We perform our GCDD algorithm (see subsection 3.1) on four skin lesion datasets. Specifically, we train ResNet50 on *ISIC2016* [13], *2017* [6], and *2019-2020* [36, 38] and train VGG16 on the *SynthDerm* [12] dataset where it contains a confirmed confounding factors "surgical markings". We summarized the discovered confounding concepts from all four datasets in Table 1. The ResNet50 trained on *ISIC2019-2020* can achieve 86.36 % ROC-AUC on the *ISIC2020* testing set, and VGG16 can achieve 100% accuracy on SynthDerm. However, by visualizing the clusters discovered by GCDD on *ISIC2019-2020* (see Fig. 5 left), it shows that ResNet50 predicts melanoma often based on confounding concepts, including dark corners, dark borders, hairs, and air pockets, while it predicts benign often based on dark borders or rulers. As for GCDD on SynthDerm, in the right of Fig. 5, it shows that VGG16 highly relies on surgical markings when predicting melanoma. More experimental results and visualization are provided in the supplementary material.

Table 1. The discovered confounding concepts from different skin lesion datasets. DC denotes dark corners, DB denotes dark borders, RL denotes rulers, APs denotes air pockets, and SMs denotes surgical markings.

Skin Lesion Dataset	DC	DB	RL	HR	APs	SMs
ISIC2019-2020	✓	✓	✓	✓	✓	
ISIC2016	✓	✓				
ISIC2017	✓	✓		✓		
SynthDerm						✓

5.3. Rewriting Model's Decision in ConfDerm

In this section, we perform experiments on our designed *ConfDerm* dataset, based on *ISIC2019-2020*, to verify whether our interaction (XIL) method can remove the confounding concept for each sub-dataset. We collect 12 clinical concepts from *Derm7pt* [17] probe dataset and five confounding concepts from the clusters produced on *ISIC2019-2020* (see Fig. 5 left). For each concept, we collect 70 positive samples and 70 negative samples to learn CAVs using a linear SVM. We define one global rule for each sub-dataset to ignore a specific confounded concept like "Do not focus on the dark corners, dark borders, rulers, hairs, and air pockets" for "*melanoma (dark corners)*", "*be-*

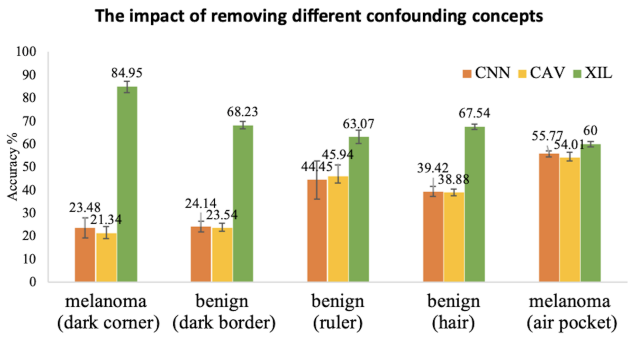


Figure 6. Performance improvement on confounded class when removing different confounded classes using XIL in five confounded datasets in the *ConfDerm* dataset.

nign (dark borders)", "*benign (rulers)*", "*benign (hairs)*" and "*melanoma (air pockets)* sub-datasets respectively." We compare our interaction method "XIL" with the CNN baseline and CNN with concept mapping (see "Turn any black box into explainable models" in section 3.3). The CNN baseline is a ResNet50 model, CAV is the CNN baseline with concept mapping, and XIL is CAV with our logical layer using human interaction learning. The impact of removing each confounding concept on the confounded class is shown in Fig. 6. For CAV, it shows that it can achieve comparable performance with CNN but cannot make the model robust to different confounding concepts. For our XIL method, the consistently better performance against CNN and CAV across all five sub-datasets demonstrates our XIL method's effectiveness in removing the model's confounding behavior. Also, it can be seen that XIL improves most on the *melanoma (dark corners)* datasets; removing dark corners can improve the accuracy of CNN from 23.48% to 84.95% for the melanoma class. This inspires us dark corners may be the main artifacts to limit the accuracy of skin cancer diagnosis, and the quality of the skin lesion dataset can be improved by avoiding collecting images with dark corners. Also, XIL gets the minimum improvement on *melanoma (air pockets)*; removing air pockets from melanoma only improves the accuracy from 55.77 % to 60 % for the melanoma class, and it may be the air pockets concept is relatively not a significant confounding factor for skin cancer diagnosis. Besides performance improvement, the correct decision-making behind the model is also essential for trustable diagnosis. In Fig. 7, both concept activation and logic explanation of our method show that the model does not perform its prediction based on the dark corners anymore after using interaction (we provide more examples in supplementary material). Moreover, we report the performance of the three methods for all classes on *ConfDerm* with five random seeds in Table 2. It can be seen that XIL improves the performance for all classes on all five sub-datasets. These results demonstrate that human interaction can help models perform better on testing sets with different distributions.

Table 2. Performance on all five sub-datasets of *ConfDerm*. MEL and BEN denote melanoma and benign. DC denotes dark corners, DB denotes dark borders, RL denotes rulers, HR denotes hairs, and APs denotes air pockets.

Datasets	CNN	CAV	XIL
MEL (DC)	60.86 ± 0.81	59.40 ± 0.93	$\mathbf{83.16 \pm 2.51}$
BEN (DB)	72.13 ± 2.05	73.40 ± 1.63	$\mathbf{75.88 \pm 1.39}$
BEN (RL)	77.87 ± 6.25	77.93 ± 5.31	$\mathbf{80.38 \pm 6.80}$
BEN(HR)	64.33 ± 2.11	64.25 ± 2.74	$\mathbf{74.91 \pm 1.52}$
MEL (APs)	60.58 ± 0.68	60.44 ± 0.95	$\mathbf{62.17 \pm 0.60}$

5.4. Debiasing the Negative Impact of Skin Tone

As illustrated in [7], existing dermatology AI models perform worse on dark skin tones as dark skin images are under-represented in most popular dermatology datasets. In this experiment, we use our XIL method again to verify whether it can reduce the negative impact of skin tone. Specifically, we use the *skincon* [8] dataset as our probe dataset to collect 22 clinical concepts and one confounding concept "dark skin". For each concept, we collect 50 positive samples and 50 negative samples and train each concept using a linear SVM. We fine-tune the DeepDerm model on the *Fitzpatrick 17K* dataset (2168 images of dark skin types compared with 7755 images of the light skin type and 6098 images of the middle skin types) and then remove the dark skin concept using XIL. In Table 3, we compare our model with the DeepDerm without XIL on the *DDI* dataset with Fitzpatrick skin tone scale labels. The consistently better performance of our method demonstrates that removing the dark skin tone concept can reduce the negative impact of skin tone and improve the robustness of the model across different skin tone type images. Further, our method achieve the most improvement on FST(V-VI), which corresponds to dark skin tone images. More results and visualizations are provided in the supplementary material.

5.5. Ablation Study

We conduct experiments to analyze the impact of key hyper-parameters and components in our framework. For the impact of the used number of concepts, we analyze it on the *Derm7pt* dataset. For other ablation studies, we perform experiments on the "*melanoma (dark corners)* dataset of ConfDerm, and we report the accuracy for all classes.

Impact of the used concept samples: We perform experiments to analyze how many concept samples are suitable for learning a high-quality concept bank. We used 30 samples as the testing set and repeatedly ran the concept learning ten times with random seeds. The mean concept accuracy is shown on the left of Fig. 8, and it can be seen that the performance saturates when using around 75 samples.

Impact of λ_1 and λ_2: The λ_1 and λ_2 in Eq. 1 are two im-

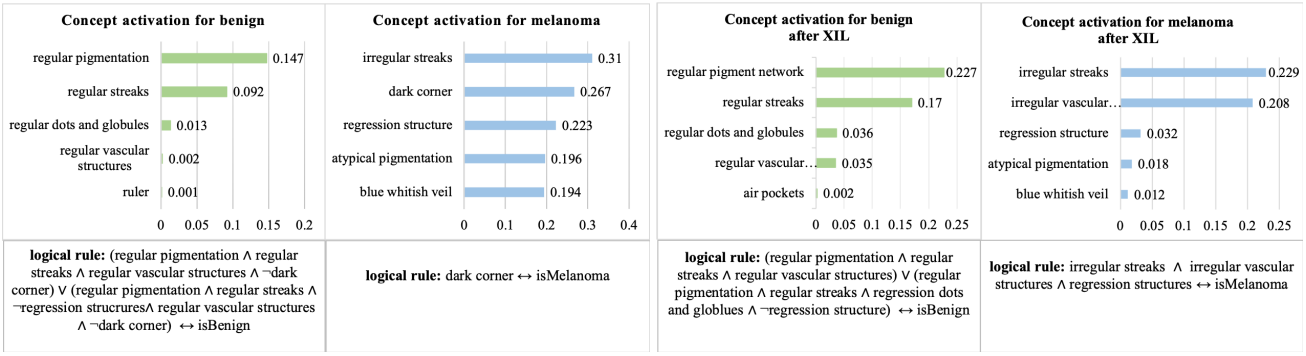


Figure 7. The global explanation of the model's behavior on the *melanoma (dark corners)* dataset of ConfDerm. In the left two figures, either the concept activation or logical rule shows that the model is confounded by the concept of the dark corners when predicting melanoma. In the right two figures, after the interaction, the model does not predict melanoma based on the dark corners, and it predicts melanoma based on meaningful clinical concepts (Zoom for better visualization).

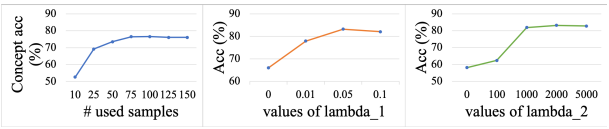


Figure 8. From left to right, the ablation on the number of used positive concept samples, the value of λ_1 of elastic-net regularization, and the value of λ_2 of the RRR loss.

Table 3. Performance (ROC-AUC) improvement using XIL on dark skin concept on the *DDI* dataset. FST (I-II) corresponds to light skin tones, FST (III-IV) corresponds to middle skin tones, and *DDI* is the combination of FST (I-VI).

Method	DDI	FST (I-II)	FST (III-IV)	FST(V-VI)
CNN	63.36 ± 0.28	64.13 ± 1.53	68.43 ± 0.70	57.933 ± 1.47
CAV	63.71 ± 0.34	64.34 ± 1.93	67.92 ± 0.73	57.22 ± 2.13
XIL	$\mathbf{64.57 \pm 0.57}$	$\mathbf{65.32 \pm 0.54}$	$\mathbf{68.77 \pm 1.17}$	$\mathbf{60.03 \pm 1.79}$

portant hyper-parameters to control the contribution of sparsity and human interaction strength. By seeing the middle of Fig. 8, we can see a relatively larger sparsity strength is helpful to improve performance as higher sparsity can make interaction easier. The right of Fig. 8 shows that setting a λ_2 with 1000 is enough for us to perform interaction to the model.

Impact of different interaction layers: In Table 4, we compare the performance when using different interaction layers, including the simple linear layer (posthoc-hoc CBM [41]), two-layer MLP, and the logic layer. We can see that two-layer MLP can achieve comparable performance with the logic layer, but the two-layer MLP cannot generate concept activations and is unable to model the relationship of concepts.

Impact of different interaction strategies: We compare different interaction strategies, including editing with normalization, L1 loss, and RRR loss in Table 4. It can be seen that editing with normalization is ineffective for the "*melanoma (dark corners)*" dataset, as editing with normalization will also hurt the performance of the unconfounded class (benign in this case). And we can see that L1 and RRR loss can alleviate this problem and achieve much better performance. The best performance of RRR loss demonstrates its effectiveness on the task.

Table 4. Comparison of the different interaction layers and interaction strategies (Accuracy for all classes).

	Editing w/ norm	L1 loss	RRR loss
Linear [41]	57.21 ± 0.03	54.30 ± 0.47	47.09 ± 0.39
MLP	-	79.42 ± 1.53	83.09 ± 3.48
Logic Layer	54.94 ± 0.19	79.17 ± 0.14	$\mathbf{83.16 \pm 2.51}$

6. Conclusion

In this paper, we show that confounding behaviors are common but rarely explored problems in skin cancer diagnosis. To tackle it, we propose a human-in-the-loop framework to effectively discover and remove the confounding behaviors of the model within the dataset during the skin cancer diagnosis. Moreover, to promote the development of skin cancer diagnosis, we crafted a new confounded dataset *ConfDerm* with different distributions between the training and testing sets. Experimental results on different datasets demonstrate our method can improve the model's performance and trustworthiness in different testing distributions.

Acknowledgment. This research was supported in part by NHMRC CRE Skin Imaging and Precision Diagnosis (APP2006551), NHMRC Synergy Roadmap Options for Melanoma Screening in Australia (APP2009923), and Airdoc-Monash philanthropic Research funding.

References

[1] Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elena Sammarco, and Mario Delfino. 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, 5

[2] Giuseppe Argenziano, Iris Zalaudek, and H. Peter Soyer. Which is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology? *British Journal of Dermatology*, 151, 2004. 2, 5

[3] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Pietro Lió, Marco Gori, and Stefano Melacci. Entropy-based logic explanations of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6046–6054, 2022. 3, 5

[4] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (De)Constructing bias on skin lesion datasets. In *ISIC Skin Image Anaylsis Workshop, 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 1

[5] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986. 4

[6] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 6

[7] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022. 5, 7

[8] Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto A Novoa, and James Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4, 7

[9] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 3, 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5

[11] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin M. Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017. 1, 2, 5

[12] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W Picard. Dissect: Disentangled simultaneous explanations via concept traversals. *arXiv preprint arXiv:2105.15164*, 2021. 6

[13] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 6

[14] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 5

[15] Jay Heo, Junhyeon Park, Hyewon Jeong, Kwang Joon Kim, Juho Lee, Eunho Yang, and Sung Ju Hwang. Cost-effective interactive attention learning with neural attention processes. In *International Conference on Machine Learning*, pages 4228–4238. PMLR, 2020. 3

[16] Sarthak Jain and Byron C. Wallace. Attention is not explanation. *NAACL 2019*, abs/1902.10186, 2019. 3, 5

[17] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. 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. 4, 6

[18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018. 2, 3, 4

[19] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 2, 3

[20] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019. 2, 4

[21] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–10. IEEE, 2020. 3

[22] Agnieszka Mikołajczyk, Sylwia Majchrowska, and Sandra Carrasco Limeros. The (de) biasing effect of gan-based augmentation methods on skin lesion images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 437–447. Springer, 2022. 1

[23] Kewen Peng and Tim Menzies. Documenting evidence of a reuse of '"why should i trust you?": explaining the predictions of any classifier'. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1600–1600, 2021. 2

[24] Samuel William Pewton and Moi Hoon Yap. Dark corner on skin lesion image dataset: Does it matter? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4831–4839, June 2022. 1

[25] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020. 6

[26] Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8116–8126. PMLR, 2020. 3

[27] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[28] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2662–2670. ijcai.org, 2017. 2, 5

[29] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.*, 2(8):476–486, 2020. 2, 4, 6

[30] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.*, 2(8):476–486, 2020. 3

[31] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2):336–359, 2020. 2, 4

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. 5

[33] Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3619–3629. Computer Vision Foundation / IEEE, 2021. 2, 3, 6

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 2

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 1, 5

[36] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 2018. 1, 5, 6

[37] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 4

[38] Rotemberg Veronica, Kurtansky Nicholas, Betz-Stablein Brigid, Caffery Liam, Chousakos Emmanouil, Codella Noel, Combalia Marc, Stephen Dusza, Guitera Pascale, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1), 2021. 1, 5, 6

[39] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. 4

[40] Fei Wang, Lawrence Casalino, and Dhruv Khullar. Deep learning in medicine—promise, progress, and challenges. *JAMA Internal Medicine*, 179, 12 2018. 1

[41] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. 2, 3, 8

[42] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2