Towards Trustable Skin Cancer Diagnosis via Rewriting Model’s Decision

Siyuan Yan\textsuperscript{1,2} Zhen Yu\textsuperscript{1,2} Xuelin Zhang\textsuperscript{1,2} Dwarikanath Mahapatra\textsuperscript{4} Shekhar S. Chandra\textsuperscript{3} Monika Janda\textsuperscript{3} Peter Soyer\textsuperscript{3} Zongyuan Ge\textsuperscript{1,2}  
\textsuperscript{1} Monash University  \textsuperscript{2} Monash Medical AI Group \textsuperscript{3} The University of Queensland \textsuperscript{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.

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 \textit{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.

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]. 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
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 \( D = \{ (x_i, y_i) \}_{i=1}^{N} \) 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 \( D \) that can confound \( f \), and then remove them via XIL on the ex-
plausible version of $f$.

The overall pipeline of our method is summarized in Fig. 3. First, we apply spectral clustering [39] on the Grad-CAM [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 $C_{confs}$ 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 $X$ from $X$. Then, we calculate the GradCAM heatmaps of images from $X$ for each class. Instead of only performing preprocessing 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 $\phi^c$ 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.

### 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 $l(h^c) = \hat{y}$ 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 $h^c$, and can be regarded as a first-order description of model’s behavior around $h^c$. 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:

$$
\mathcal{L}(h^c, y, H) = \mathcal{L}_{CE}(y, \hat{y}) + \lambda_1 \left( \alpha \| w \|^1 + (1 - \alpha) \| w \|^2 \right) + \lambda_2 \sum_{b=1}^{B} \sum_{d=1}^{D} \left( \frac{H_{bd}}{1 + \left( \frac{H_{bd}}{\sum_c (\hat{y}_{bc}}) \right)^2} \right) \tag{1}
$$

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 $\lambda_1$ controls the sparsity of $\hat{l}$, and $\lambda_2$ 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.

**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-
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.

### 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-
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 confounded 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 trustworthy 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.

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 775 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 DD1 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 types. 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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>CNN</th>
<th>CAV</th>
<th>XIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEL (DC)</td>
<td>60.86 ± 0.81</td>
<td>59.40 ± 0.93</td>
<td>83.16 ± 2.51</td>
</tr>
<tr>
<td>BEN (DB)</td>
<td>72.13 ± 2.05</td>
<td>73.40 ± 1.63</td>
<td>75.88 ± 1.39</td>
</tr>
<tr>
<td>BEN (RL)</td>
<td>77.87 ± 6.25</td>
<td>77.93 ± 5.31</td>
<td>80.38 ± 6.80</td>
</tr>
<tr>
<td>BEN (HR)</td>
<td>64.33 ± 2.11</td>
<td>64.25 ± 2.74</td>
<td>74.91 ± 1.52</td>
</tr>
<tr>
<td>MEL (APs)</td>
<td>60.58 ± 0.68</td>
<td>60.44 ± 0.95</td>
<td>62.17 ± 0.60</td>
</tr>
</tbody>
</table>

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.
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 \( \lambda_1 \) of elastic-net regularization, and the value of \( \lambda_2 \) 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).

<table>
<thead>
<tr>
<th>Method</th>
<th>DDI</th>
<th>FST (I-II)</th>
<th>FST (III-IV)</th>
<th>FST (V-VI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>63.36 ± 0.28</td>
<td>64.13 ± 1.53</td>
<td>68.43 ± 0.70</td>
<td>57.933 ± 1.47</td>
</tr>
<tr>
<td>CAV</td>
<td>63.71 ± 0.34</td>
<td>64.34 ± 1.93</td>
<td>67.92 ± 0.73</td>
<td>57.22 ± 2.13</td>
</tr>
<tr>
<td>XIL</td>
<td>64.57 ± 0.57</td>
<td>65.32 ± 0.54</td>
<td>68.77 ± 1.17</td>
<td>60.03 ± 1.79</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Editing w/ norm</th>
<th>L1 loss</th>
<th>RRR loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear [41]</td>
<td>57.21 ± 0.03</td>
<td>54.30 ± 0.47</td>
<td>47.09 ± 0.39</td>
</tr>
<tr>
<td>MLP</td>
<td>-</td>
<td>79.42 ± 1.53</td>
<td>83.09 ± 3.48</td>
</tr>
<tr>
<td>Logic Layer</td>
<td>54.94 ± 0.19</td>
<td>79.17 ± 0.14</td>
<td>83.16 ± 2.51</td>
</tr>
</tbody>
</table>

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.

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


