Towards Trustable Skin Cancer Diagnosis via Rewriting Model’s Decision
–Supplemental Material–

Siyuan Yan, Zhen Yu, Xuelin Zhang, Dwarikanath Mahapatra,
Shekhar S. Chandra, Monika Janda, Peter Soyer, Zongyuan Ge

Monash University, Monash Medical AI Group, The University of Queensland,
Inception Institute of AI, Abu Dhabi, UAE

Abstract

In this supplementary material, we provide more details about datasets, additional training details, network architectures, t-SNE visualisation, concept accuracy, and explanation visualisation.

1. More Details about Datasets

1.1. General Datasets

We choose SynthDerm, ISIC2016, ISIC2017, and ISIC2019_2020 as evaluating datasets in the section “Confounding Concept Discovery” of the experimental part. We chose SynthDerm as it is a well-controlled dataset and chose other datasets due to their popularity in dermatology. Also, we choose Fitzpatrick17k and DDI as training and testing dataset in the section “Debiasing the Negative Impact of Skin Tone” as they contain rich Fitzpatrick skin type labels.

SynthDerm: SynthDerm [10] is a balanced synthetic dataset inspired by real-world ABCD rule criteria [2] of melanoma skin lesions. It includes images with different factors, including whether asymmetric, different borders, colors, diameter, or evolving in size, shape, and color over time. For skin tone, it simulates six Fitzpatrick skin scales. It includes 2600 64x64 images. Moreover, in this dataset, there are surgical markings in melanoma images but not in benign images. Thus, the “surgical markings” is the confounding factors in the dataset.

ISIC2016: We use the data from the task 3 of ISIC2016 [12] challenge, it contains 900 dermoscopic images.


DDI: DDI [6] is similar to Fitzpatrick17k but with higher quality. It contains 208 images of FST (I-II), 241 images of FST (III-IV), and 207 images of FST (I-VI), which corresponds to light skin, middle skin, and dark skin tone, respectively.

1.2. Probe Datasets:

For constructing the concept bank, we use Derm7pt as the probe dataset for dermoscopic image dataset such as ConfDerm and use SKINCON as the probe dataset for clinical image dataset such as Fitzpatrick17k.

SKINCON: SKINCON [7] is a skin disease dataset densely annotated by domain experts for fine-grained model debugging and analysis. It includes 3230 images with 48 clinical concepts, 22 of which have over 50 images.

Derm7pt: Derm7pt [14] is a dermoscopic image dataset contains 1011 dermoscopic images with 7 clinical concepts (i.e., pigmentation network, blue whitish veil, vascular structures, pigmentation, streaks, dots and globules, and regression structures.) [1] for melanoma skin lesions in dermatology.

1.3. ConfDerm:

We provide additional data visualization, showing the characteristics of images in the confounded class of five datasets in our ConfDerm dataset, as illustrated in Fig. 1.

2. Additional Training Details and Network Architecture

Detail of the logic layer: We choose the recently proposed entropy-based logical layer [3]. It consists of four steps: (1) For each concept, calculate the concept importance score $\gamma_j$ via calculating the $l^2$ norm of all neurons in subsequent layers connected to the concept. (2) Perform softmax and rescaling on the $\gamma$. (3) Get the importance-aware concept score $h^c$ via weighting the $\gamma$ on all concept scores $h^c$. (4) Finally, feed the $h^c$ into subsequent layers. The first-order logic generation of the model is described in the example
of Fig. 2. It binaries the concept scores $h^c$ and the attention weights $\gamma_j$, then select one concept if its weight $\gamma_j$ is 1.

This method is based on attention operation, but [8, 13] shows that attention is often not the explanation, which causes interaction on it is not effective in changing the model’s behavior. In Fig. 2, it shows that global explanations of the model using attention and our method, after the interaction, the left of Fig. 2 shows that the model using attention still focuses on the "ruler" concept, and the right of Fig. 2 shows that the model using our explanation does not give a high weight for the ruler and can focus on meaningful clinical concepts.

Training Details for "Rewriting Model’s Decision in ConfDerm" : For concept bank construction, we train a linear SVM using the sklearn library [4] with regularization $\beta = 0.14$ for each concept. We totally train 17 concept vectors, where 12 concepts are from the Derm7pt dataset and 5 concepts from our GCCD algorithm. All 17 concepts we obtained are "regular_pigment_network", "irregular_pigment_network", "blue_whitish_veil", "regular_vascular_structures", "melanoma (dark corners)", "benign (dark borders)", "benign (rulers)", "benign (hairs)", "melanoma (air pockets)".

Figure 1. Visualization of the confounded class of five datasets in our ConfDerm, each one has one confounding factor, including dark corners, dark borders, hairs, and air pockets.

For model training, we train our framework using PyTorch with a maximum of 20 epochs on each subdataset on ConfDerm dataset. Each image is rescaled to $256 \times 256$. The black-box model is initialized with ResNet50 trained on ImageNet, and we set the logic layer using two linear layers. We use Adam optimiser and set the learning rate with $0.001$, and we set the balanced weights $\lambda_1$ and $\lambda_2$ of our loss with $0.05$ and $2000$.

Training Details for "Debiasing the Negative Impact of Skin Tone": For concept bank construction, similarly, we train a linear SVM using the sklearn library \cite{4} with regularization $\beta = 0.1$ for each concept. We choose 22 concepts that have at least 50 images and one additional confounding concept, "dark skin" from the SKINCON dataset. To the end, all 23 concepts we collected are "Papule", "Plaque", "Pustule", "Bulla", "Patch", "Nodule", "Ulcer", "Crust", "Erosion", "Atrophy", "Exudate", "Telangiectasia/Scale", "Scar", "Friable", "Dome-shaped", "Brown(Hyperpigmentation)", "White(Hypopigmentation)", "Purple", "Yellow", "Black", "Erythema", "dark skin".

For model training, we split the Fitzpatrick17k dataset into training and validation set with a ratio of 8:2 and use DDI dataset as the testing set. We train our framework using PyTorch with a maximum of 30 epochs and use Adam optimiser, and set the learning rate with $3e^{-4}$, and we set the balanced weights $\lambda_1$ and $\lambda_2$ of our loss with 0.1 and 4000. Each image is rescaled to $256 \times 256$. The black-box model is initialized with InceptionV3 \cite{15} trained on the dataset \cite{9}, as similar to \cite{7}, and we set the logic layer using three linear layers.

3. Additional Experiments

3.1. More Visualisation about Confounding Concept Discovery

GCDD on ISIC2016 and ISIC2017: We also visualize the t-SNE of our GCDD algorithm within ISIC2016 and ISIC2017, as shown in Fig. 4.

Samples of Representative Clusters within ISIC2019_2020: The representative clusters of GCDD on ISIC2019_2020 are illustrated in Fig. 5.

3.2. Concept Learning

We report the testing accuracy of each concept in Derm7pt and SKINCON dataset, as shown in Table 1 and Table 2.
3.3. More Analysis about Global Explanations

Explanations of “Rewriting Model’s Decision in Conf-Derm”: We provide the comparison between the explanation of the model and the explanation of the model after XIL on other four datasets, including benign (dark borders), benign (rulers), benign (hairs), and melanoma (air pockets), as shown in Fig. 6, 7, 8, and 9. It can be seen that our XIL method can make the model focus less on confounding factors.

Explanations of “Debiasing the Negative Impact of Skin Tone”: In Fig. 10, we show the comparison between the explanation of the model and the explanation of the model after XIL on Fitzpatrick17k dataset. It can be seen that our XIL method makes the model focus less on dark skin and can focus on meaningful clinical concepts again.
Figure 6. The global explanation of the model’s behavior on the benign (dark borders) dataset of ConfDerm. In the left figure, either the concept activation or logical rule shows that the model is confounded by the concept of the "dark border" when predicting benign. In the right figure, after the interaction, the model does not predict benign based on the dark corners, and it predicts benign based on meaningful clinical concepts.

Figure 7. The global explanation of the model’s behavior on the benign (rulers) dataset of ConfDerm. In the left figure, either the concept activation or logical rule shows that the model is confounded by the concept of the "ruler" when predicting benign. In the right figure, after the interaction, the model relies less on "ruler" and can predict benign based on meaningful clinical concepts.

Figure 8. The global explanation of the model’s behavior on the benign (hairs) dataset of ConfDerm. In the left figure, either the concept activation or logical rule shows that the model is confounded by the concept of the "hair" when predicting benign. In the right figure, after the interaction, the model relies much less on "hairs" and can predict benign based on meaningful clinical concepts.

Table 2. Concept accuracy on testing set of SKINCON.

<table>
<thead>
<tr>
<th>Concept name</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papule</td>
<td>65</td>
</tr>
<tr>
<td>Plaque</td>
<td>72.5</td>
</tr>
<tr>
<td>Pustule</td>
<td>81.82</td>
</tr>
<tr>
<td>Bulla</td>
<td>82.57</td>
</tr>
<tr>
<td>Patch</td>
<td>66.67</td>
</tr>
<tr>
<td>Nodule</td>
<td>76.32</td>
</tr>
<tr>
<td>Ulcer</td>
<td>84.38</td>
</tr>
<tr>
<td>Crust</td>
<td>60</td>
</tr>
<tr>
<td>Erosion</td>
<td>72.5</td>
</tr>
<tr>
<td>Atrophy</td>
<td>57.14</td>
</tr>
<tr>
<td>Exudate</td>
<td>86.67</td>
</tr>
<tr>
<td>Telangiectasia</td>
<td>80</td>
</tr>
<tr>
<td>Scale</td>
<td>73.89</td>
</tr>
<tr>
<td>Scar</td>
<td>65.38</td>
</tr>
<tr>
<td>Friable</td>
<td>83.33</td>
</tr>
<tr>
<td>Dome-shaped</td>
<td>70</td>
</tr>
<tr>
<td>Brown (Hyperpigmentation)</td>
<td>65</td>
</tr>
<tr>
<td>White (Hypopigmentation)</td>
<td>50</td>
</tr>
<tr>
<td>Purple</td>
<td>66.67</td>
</tr>
<tr>
<td>Yellow</td>
<td>67.5</td>
</tr>
<tr>
<td>Black</td>
<td>83.33</td>
</tr>
<tr>
<td>Erythema</td>
<td>77.5</td>
</tr>
<tr>
<td>dark skin</td>
<td>80</td>
</tr>
</tbody>
</table>

Figure 9. The global explanation of the model’s behavior on the melanoma (air pockets) dataset of ConfDerm. In the left figure, either the concept activation or logical rule shows that the model is confounded by the concept of the "air pockets" when predicting melanoma. In the right figure, after the interaction, the model does not predict melanoma based on “air pockets” and can predict benign based on meaningful clinical concepts.
Figure 10. The global explanation of the model’s behavior on the Fitzpatrick17k dataset. In the two left figures, either the concept activation or logical rule shows that the model is confounded by the concept of the dark corners when predicting malignant. In the two right figures, after the interaction, the model relies less on “dark skin” to predict malignant, and it predicts malignant based on meaningful clinical concepts.
References


