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

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

*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.

**ISIC2017:** We use the data from the part 3 of *ISIC2017* [5] challenge, it contains 2000 dermoscopic images.

**ISIC2019_2020:** *ISIC2019_2020* [16, 17] is the *ISIC2020* dataset with all melanoma images from *ISIC2019*, which includes 37648 dermoscopic images.

**Fitzpatrick17k:** *Fitzpatrick17k* [11] contains 16577 clinical images labeled by 114 skin conditions and 6 Fitzpatrick skin types.

**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 $l2$ 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 $\hat{h^c}$ via weighting the $\gamma$ on all concept scores $h^c$. (4) Finally, feed the $\hat{h^c}$ into subsequent layers. The first-order logic generation of the model is described in the example

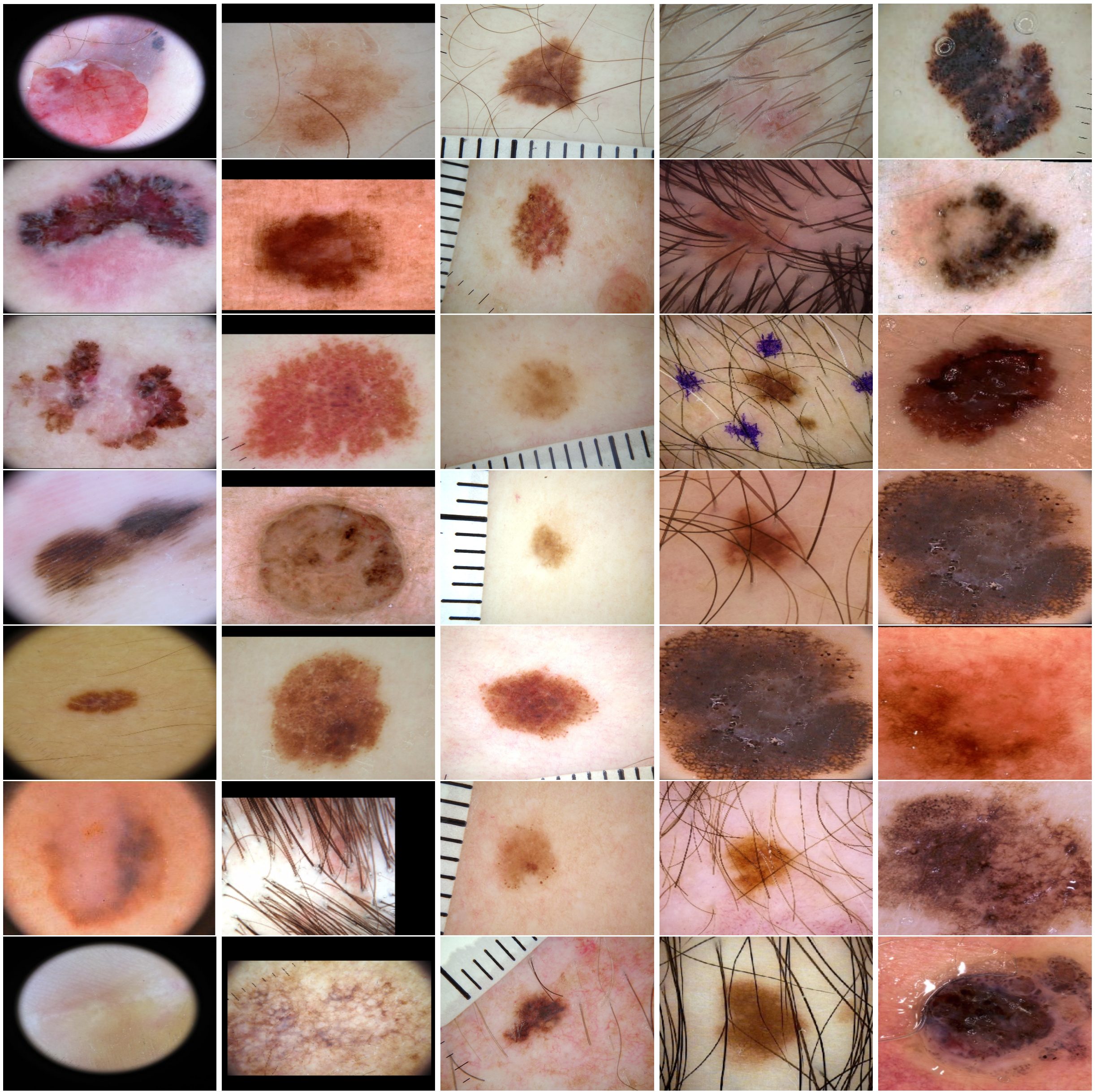

Figure 1. Visualization of the confounded class of five datasets in our *ConfDerm*, each one has one confounding factor, including dark corners, dark borders, hairs, and air pockets.

of Fig. 2. It binaries the concept scores $h^c$ and the attention weights $\gamma$, 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",*

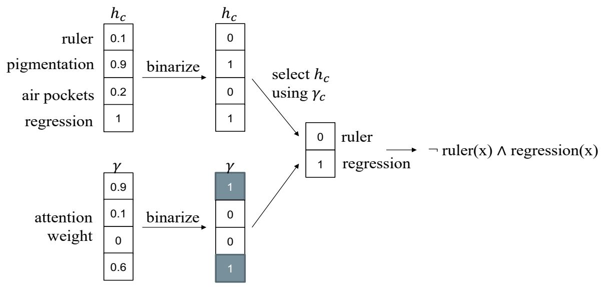

Figure 2. Illustration of the logical explanation generation.

*“irregular_vascular_structures", “typicalpigmentation", “atypical_pigmentation", “regular_streaks", “irregular_streaks", “regular_dots_and_globules", “irregular_dots_and_globules", “regression_structures", “dark corner", “dark border", “air pockets", “ruler", “hair".*

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 × 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 [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 × 256. The black-box model is initialized with InceptionV3 [15] trained on the dataset [9]. as similar to [7], and we set the logic layer using three linear layers.

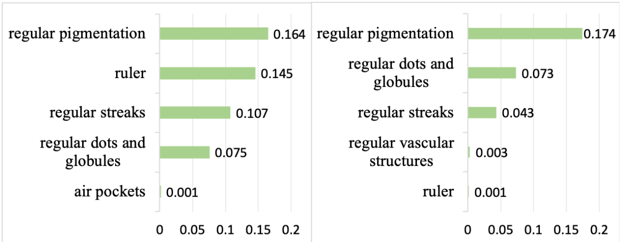

Figure 3. Global explanation (concept activation) of reducing the confounding factor "rulers" on *ConfDerm* dataset. From left to right, Interaction on attention, Interaction on our explanation. Results show that interaction with our explanation successfully alleviate the negative impact of the confounding factor.

Table 1. Concept accuracy on testing set of Derm7pt.

| concept name | Acc (%) |
|---|---|
| regular_pigment_network | 77.5 |
| irregular_pigment_network | 72.5 |
| blue_whitish_veil | 70 |
| regular_vascular_structures | 70 |
| irregular_vascular_structures | 63.33 |
| typicalpigmentation | 77.5 |
| atypical_pigmentation | 65 |
| regular_streaks | 67.5 |
| irregular_streaks | 62.5 |
| regular_dots_and_globules | 67.5 |
| irregular_dots_and_globules | 72.5 |
| regression_structures | 70 |
| dark corner | 100 |
| dark border | 100 |
| air pockets | 100 |
| ruler | 100 |
| hair | 100 |
| dark skin | 85 |

# 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.

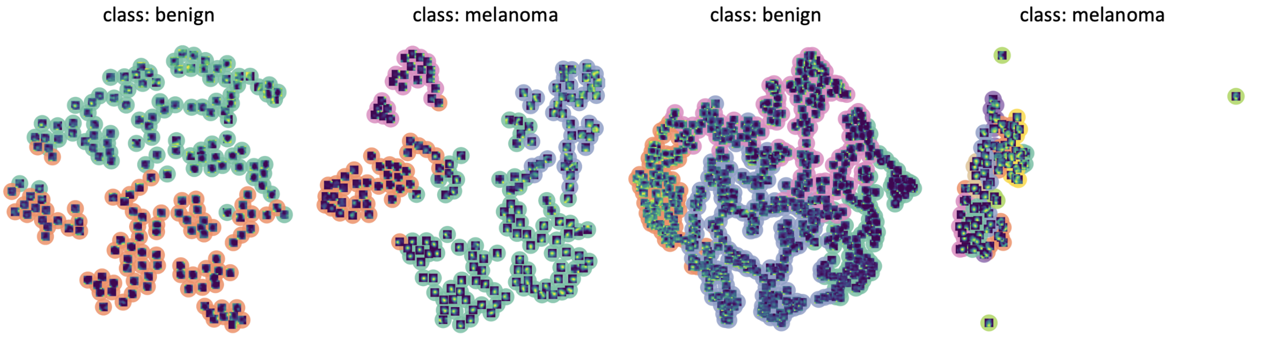

Figure 4. Global analysis of the models' behavior within datasets using GCCD. The two left graphs are the tSNE of spectral clustering using GradCAMs of a ResNet50 within *ISIC2016*. The two right are the tSNE of spectral clustering using GradCAMs of a ResNet50 within the *ISIC2017* dataset.

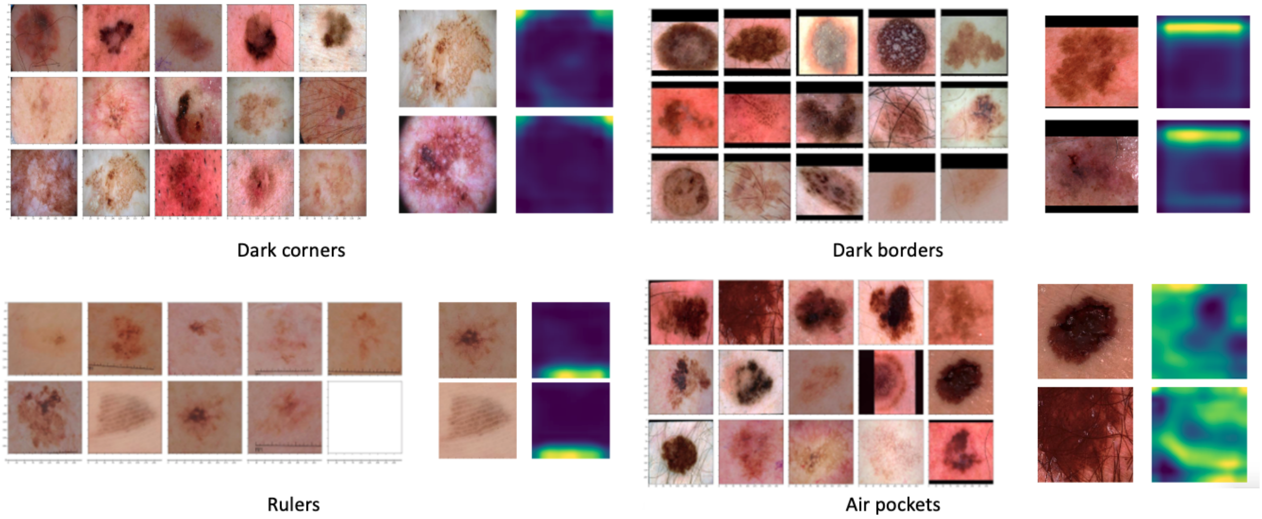

Figure 5. Visulisation the representative clusters of GCCD on *ISIC2019_2020*.

### 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.

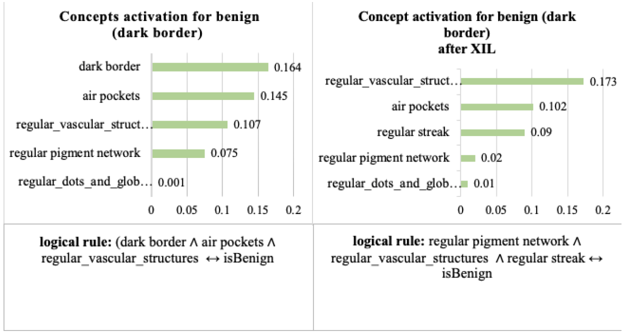

Figure 6. The global explanation of the model's behavior on the benign (dark borders) dataset of ConfDerm. In the left figure, either the concept activation or logical rule shows that the model is confounded by the concept of the "dark border" when predicting benign. In the right figure, after the interaction, the model does not predict benign based on the dark corners, and it predicts benign based on meaningful clinical concepts.

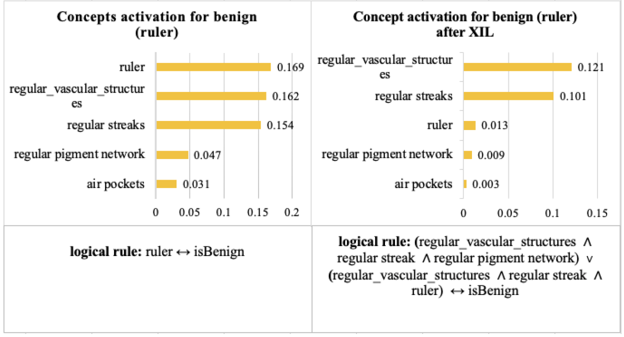

Figure 7. The global explanation of the model's behavior on the benign (rulers) dataset of *ConfDerm*. In the left figure, either the concept activation or logical rule shows that the model is confounded by the concept of the "ruler" when predicting benign. In the right figure, after the interaction,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

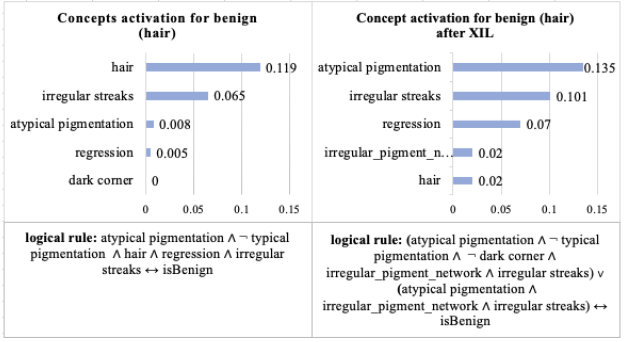

Figure 8. The global explanation of the model's behavior on the benign (hairs) dataset of *ConfDerm*. In the left figure, either the concept activation or logical rule shows that the model is confounded by the concept of the "hair" when predicting benign. In the right figure, after the interaction, the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

Table 2. Concept accuracy on testing set of SKINCON.

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

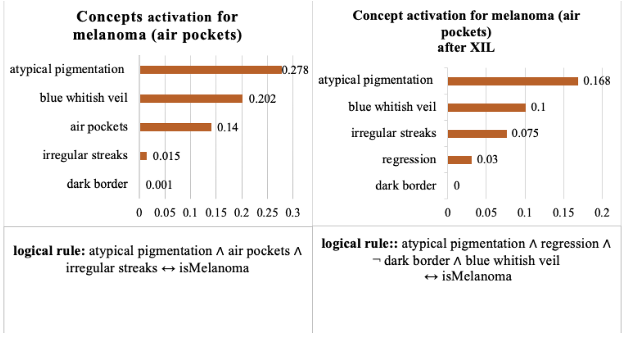

Figure 9. The global explanation of the model's behavior on the melanoma (air pockets) dataset of *ConfDerm*. In the left figure, either the concept activation or logical rule shows that the model is confounded by the concept of the "air pockets" when predicting melanoma. In the right figure, after the interaction, the model does not predict melanoma based on “air pockets" and can predict benign based on meaningful clinical concepts.

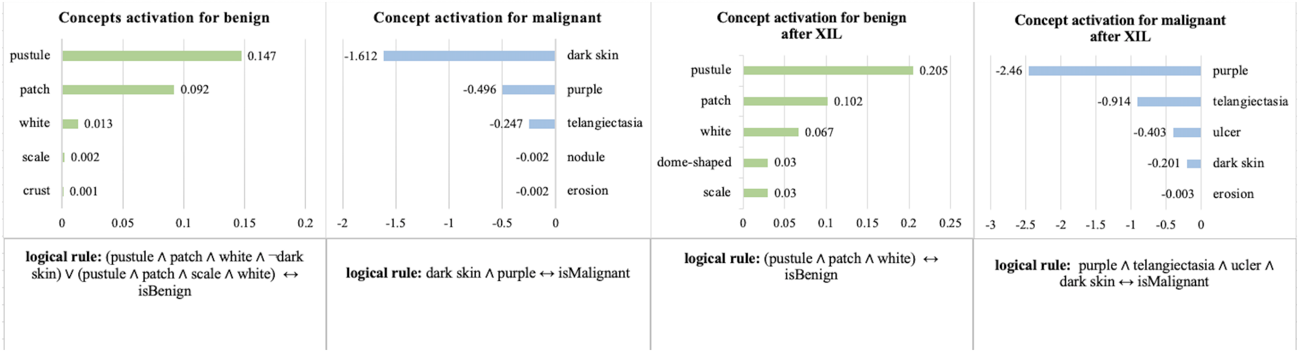

Figure 10. The global explanation of the model's behavior on the *Fitzpatrick17k* dataset. In the two left figures, either the concept activation or logical rule shows that the model is confounded by the concept of the dark corners when predicting malignant. In the two right figures, after the interaction, the model relies less on "dark skin" to predict malignant, and it predicts malignant based on meaningful clinical concepts.

## References

[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. 1

[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. 1

[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. 1

[4] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013. 2, 3

[5] 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. 1

[6] 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. 1

[7] 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. 1, 3

[8] 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. 2

[9] 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. 3

[10] 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. 1

[11] Matt Groh, Caleb Harris, Luis R. Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828, 2021. 1

[12] 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. 1

[13] Sarthak Jain and Byron C. Wallace. Attention is not explanation. *NAACL 2019*, abs/1902.10186, 2019. 2

[14] 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. 1

[15] 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. 3

[16] 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

[17] 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