Interpretable AI for Dermoscopy Images of Pigmented Skin Lesions

M. Defresne[1,2], E. Coutier[1,3], P. Fricker[1,2], F.A.A. Blok[2,4], H. Nguyen[1,2]

[1] Torus AI, 12 Av. de l'Europe, Ramonville-Saint-Agne, France
[2] BelleTorus Corporation (Belle.ai), 245 First Street, Riverview II, 18th Floor, Cambridge, MA, USA
[3] Université de Toulouse - IMT Mines Albi, France
[4] Department of Dermatology, Amsterdam University Medical Center, The Netherlands

{marianne.defresne, hangntt}@torus-actions.fr

Résumé

L'apprentissage profond est efficace pour la détection du cancer de la peau à partir d'images de lésions. Cependant, son adoption pratique reste limitée par le manque d'explication derrière ses décisions. Notre modèle n'échappant pas à la règle, nous l'analysons à partir de deux règles de diagnostic dermatologique. Premièrement, nous proposons un outil de visualisation pour fournir aux praticiens un contexte supplémentaire à la décision de notre modèle. Deuxièmement, nous présentons une variante du modèle basée sur des concepts médicaux, qui améliore l'interprétabilité du modèle initial.

Mots-clés

IA explicable, XAI, cancer de la peau, aide à la décision

Abstract

Deep Learning is successful at detecting skin cancer from a lesion image, but its practical adoption is limited by the lack of explanation behind its decisions. The model we develop is no exception; we aim to analyze it based on two dermatology rules for lesion diagnosis. First, we propose a visualization tool to give practitioners additional context to our model's decision. Second, we introduce a model variant based on medical concepts that enhance the interpretability of the baseline model.

Keywords

Explainable AI, XAI, skin cancer, aided decision support

1 Introduction

Skin cancers, including melanoma, are one of the most common cancers in the world [19]. Early diagnosis is crucial to reduce morbidity and mortality [16]. Human diagnosis is based primarily on visual inspection, often with a dermatoscope for more details, and comprehensive rules. The ABCD rule [15] and the 7-point checklist (7PCL) [2] are the most common. The ABCD rule provides a decision based on asymmetry, border irregularity, color variation, and dermoscopic structure of the lesion, while the 7-point checklist (7PCL) provides a score based on 7 visual signs to detect suspicious lesions.

Recent advances in Deep Learning have been successfully applied to skin lesion classification, as neural networks outperform dermatologists [8]. In this trend, we developed a CNN-based tool to assist dermatologists[1][2]. For now, the model lacks interpretability and it is not yet able to give explanation to doctors to motivate their decision. This limits the trust practitioners have in Artificial Intelligence (AI) and thus its adoption. Moreover, training data itself contains biases non-meaningful for humans but exploited by classification models [3]. Understanding and quantifying how much of the decision aligns with medical concepts versus biases indicates the model's robustness. This paper aims to give insights into our neural network's behavior and provide meaningful explanations to practitioners.

Contributions

After briefly introducing our neural net for skin lesion classification, we illustrate biases in the dataset to motivate further our developments, which are summarized as follows:

1. We develop non-neural algorithms to assess the criteria of the ABCD rule as a tool for practitioners.
2. We analyze the medical concepts learned by our model using the 7PCL.
3. We transfer our model to a concept-based model to explain its decision based on medical signs.

Related works

There have been massive efforts to develop neural networks for skin cancer prediction [10], partly driven by the challenges hosted by the International Skin Imaging Collaboration (ISIC) between 2016 and 2020 [6] that crowned winners [5, 7] at each session. The ongoing IToBoS[3] European consortium aims for a Computer Aided Diagnostics tool for melanoma and emphases providing valuable explanations for practitioners.
Explainable AI (XAI) has been applied to skin cancer classification, as recently reviewed [9]. Most approaches focus on *post-hoc* explanations via heatmap [22] or feature importance [18], which requires human interpretation and

[1] https://play.google.com/store/apps/details?id=com.bellepro.app
[2] https://apps.apple.com/in/app/bellepro/id1615008664
[3] https://itobos.eu/

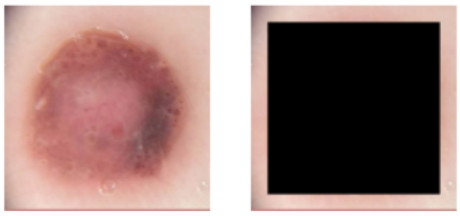

Figure 1: Left: input image. Right: image with box mask.

is prone to confirmation bias and cherry-picking. Few approaches incorporate interpretability, for instance, by predicting medical signs [11] or by retrieving visually similar images [21]. As often in XAI [13], a trade-off appears between classification accuracy and explainability.

2 Method

We first describe our skin lesion classification model and conduct a brief experiment to show that the dataset is biased. The need to enhance doctors' trust motivates us to investigate *a posteriori* how well the model decision aligns with dermatology concepts and to introduce an interpretable concept-based model.

2.1 Classification model and dataset

Our model is based on an efficientnet-b4 [20], pre-trained on ImageNet, to extract embeddings and one classification head (2 linear layers with final Sigmoid activation) to predict the skin cancer class. A focal loss is used to train the model. The model predicts 10 classes, but we focus the explanations on benign vs malignant classes.

We used an internal dataset of $104,000$ images. Inspired by a previous approach [3] revealing biases in the ISIC dataset, we investigate how biased our dataset is. Indeed, the presence of biases creates spurious correlations that the model may exploit instead of meaningful information, leading to a seemingly better accuracy but lower robustness. We reproduce part of their methodology by training and testing a new model on altered images where a box mask covering 70% of pixels is applied, hiding the entirety of the lesion, as illustrated in Figure 1.

When comparing the baseline model with the one trained with box mask (see Table 1), classification performances drop, but they are far from random, *i.e.,* predicting class based on frequency in the train set. The AUROC is even similar to the performances of dermatologists [4] (assessed on a different train set). Since the model can make better-than-random classification when the lesion is masked, we deduce the existence of spurious correlations within the dataset. This motivates the importance of knowing which information the model uses, as we investigate in the following sections.

2.2 *Post-hoc* interpretability

2.2.1 Applying the ABCD rule

The ABCD rule is a simple yet effective tool for dermatologists to assess pigmented lesions for potential malignancy [15]. Here, we apply each ABCD criterion as an image-processing technique to analyze dermoscopic images. We leverage it to provide additional visual information to clinicians, as illustrated in Figure 2. We additionally derive metrics from ABC criteria to study their correlation with our dataset's ground truth.

	Baseline	Box model	Random	Doctors
Accuracy	73%	35%	17.8%	-
AUROC	0.92	0.68	-	0.67 [4]

Table 1: Classification metrics on internal test set.

Asymmetry. Irregular shapes and uneven color distribution are strong indications of melanoma. We quantify asymmetry by first isolating the lesion using segmentation. Then, we identify potential symmetry axes based on the segmented region's shape. Crucially, these axes are informed by the prior shape asymmetry analysis, ensuring color assessment aligns with potential shape irregularities. If there is no symmetry in the shape, there will be no symmetry in colors. We calculate the Intersection over Union (IoU) between the lesion and its mirrored counterpart for each axis, considering both shape and color distribution weighted by color presence. A significant deviation from an IoU of 1 (indicating perfect overlap) suggests asymmetry, potentially signifying melanoma.

Border irregularity. Smooth borders are characteristic of benign lesions. Conversely, melanomas often exhibit notched, sharp, or uneven borders. We assess border regularity using the convex hull method. By comparing the convex hull representing the smallest convex shape encompassing the entire lesion, we highlight the discrepancy between its boundaries and its convex hull. This highlights the number of deviations and indicates the lesion irregularity and its potential malignancy.

Color variation. To analyze color variation — melanomas often exhibit a mix of brown, black, blue, white, and red hues compared to uniformly brown benign nevi — we perform image normalization. This ensures a consistent color basis across images, addressing illumination variations.

We then segment the lesion into smaller regions and compute the most frequent color in each, based on a dermatologist-defined list [17].

Differential structures. When visualizing features and patterns within a lesion through dermoscopy, dermatologists refer to differential structures — or dermoscopic structures — such as the pigment network and vascular patterns as indicators of malignancy. Atypical features like blue-white veil or irregular pigmentation are strongly associated with melanoma. We leverage the Meijering filter [14] to analyze these structures.

2.2.2 Testing learned concepts

We further analyze the concepts learned by the model using medical annotations. The Interactive Atlas of Dermoscopy [1] contains 1011 dermoscopic images with diag-

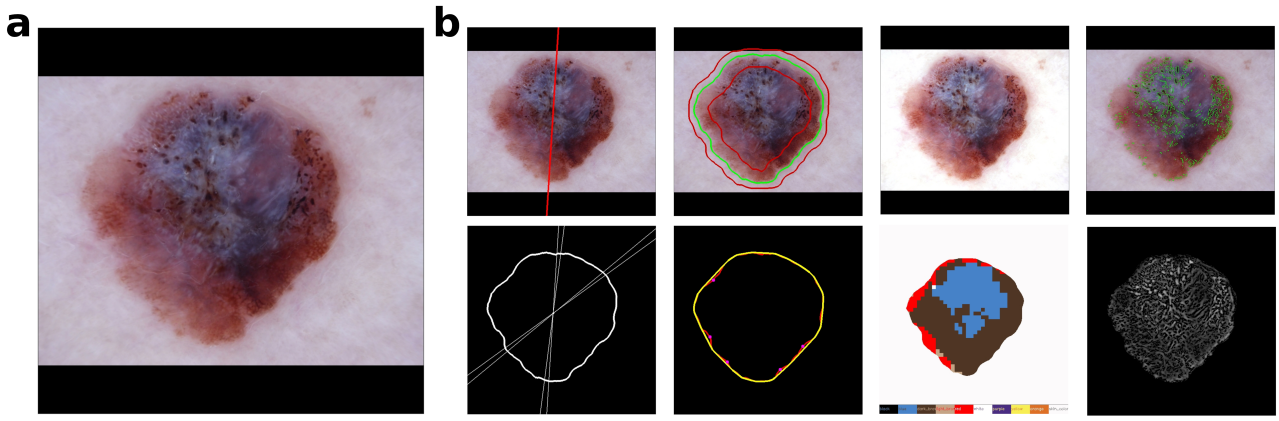

Figure 2: Applying the ABCD criteria on a dermoscopic image featuring a melanoma (best viewed in color). **a.** The original dermoscopic image pad to a square format. **b.** From left to right: the asymmetry, border, color, and dermoscopic structure criteria applied to the dermoscopic image. Top row: the best symmetry axis found based on shape and color, the border of the lesion in green with its inner and outer border in red, the normalized image, the highlighted in green dermoscopic structure. Bottom row: all the symmetry axes found with an IoU of at least 0.9, the convex hull based on the detected border of the lesion in yellow, the detected border of the lesion in red, and the deviations as pink dots, the color variations inside the lesion, the enhanced result of the Meijering filter.

nosis and a label for each sign of the 7PCL. These signs describe patterns in the lesion, such as streaks, dots and globules or pigmentation, some of them indicating a suspicious lesion. Our model never sees this dataset.

We use those signs as medical concepts and test whether our model uses them with the Testing with Concept Activation Vector (TCAV) approach [12]. We first create two banks of lesion images for each sign of a suspicious lesion: one bank containing the malignant sign and one without it (*i.e.*, sign labeled as "absent" or "regular" or "typical"). We fit a linear classifier with class weights to separate the embeddings into two classes. A concept is the vector normal to the decision boundary. We then test if this concept is important for a class (nevus or melanoma) using a third bank of images of the class, with no overlap with the other two banks. For each image, the derivative of the class logit w.r.t the embedding is computed. The TCAV score is the ratio of images whose derivative is in the same direction as the concept vector (*i.e., positive dot product)*. A score near 1 means the concept is important for the class, while a score around 0.5 corresponds to a random concept.

2.3 Interpretability by design

The TCAV method gives insight into the global model behavior but does not provide a precise explanation for a single image. Thus, we explore an alternate model that predicts the 7 signs of 7PCL instead of predicting classes. The lesion is detected as malignant if it has a score higher than 3. The decision is then made based on detected signs: a major (resp. minor) sign gives 2 (resp. 1) points, and the lesion is suspected as malignant if it has a score of 3 or more.

As the Atlas dataset is small, we do not train the new model from scratch but reuse the previous encoder and freeze its weights. We train 7 classification heads (single lineal layer with Sigmoid each), one for each sign. The loss rewards a correct classification of each sign individually (using cross-entropy) and a correct diagnosis by comparing the true score to the score computed from predicted signs. Models are trained until the mean validation accuracy increases.

3 Results

We first analyze our baseline model's decision *a posteriori* and compare it to its interpretable concept-based version.

3.1 *Post-hoc* interpretability

ABCD rule. We assess whether the criteria we derived from the ABCD rule can be used as an indicator for practitioners, along with the class prediction. Most of our criteria are visual, as illustrated in Figure 2. We add quantitative results by comparing 3 distributions for benign and malignant classes (including melanoma, basal cell carcinoma, and epidermal tumors). We consider the number of orthogonal symmetry axes (to within 20 degrees), the highest distance between the lesion's actual border and the convex hull, and the number of colors in a single lesion. As shown in Figure 3, we observe a distribution shift between benign and malignant lesions on the 3 criteria. As expected, malignant lesions tend to have fewer symmetry axes, a border further away from its convex hull, and contain more colors.

The classification uses medical concepts. We test the importance of each sign of 7PCL for the classification with TCAV, each sign corresponding to a concept. For each sign and each class, the concept is computed 10 times on random train/test splits of images with and without concept.

As displayed in Figure 4, all signs of the 7PCL are important for classifying a melanoma, with scores ~ 1. On the contrary, all signs disfavor the nevus class.

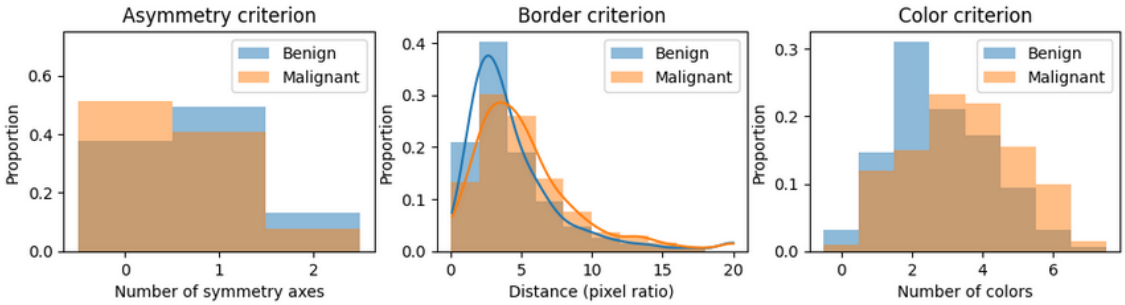

Figure 3: Assessing asymmetry (left), border irregularity (middle), and color variation (right) criteria on the test set. Brown is the overlap between two distributions.

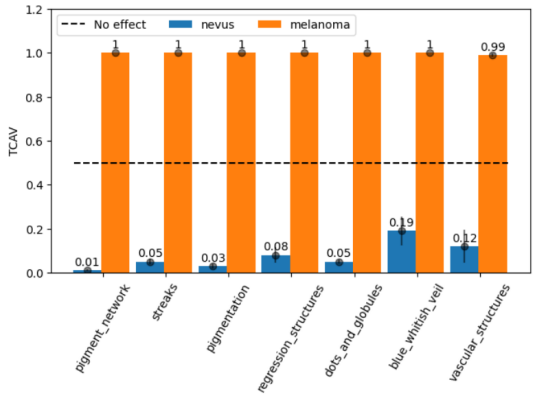

Figure 4: TCAV on the 7PCL, mean and std on 10 runs.

3.2 Interpretability by design

We first assess the interpretable model by its average binary accuracy on each sign (malignant/benign) and the mean absolute error (MAE) between the resulting score and the true score. For reference, true scores range from 0 to 7. Table 2 displays that accuracy averaged over signs is better with CE-only training, but the resulting score is further from the true score. The accuracy is surprisingly low: even if the concepts derived from these signs are important for the classification model, it cannot recognize their presence reliably. This could be due to the small size of the Atlas dataset.

Training	MSE+CE	CE-only
Mean acc.	74.4%	75.9%
MAE	1.41	1.71

Table 2: Test metrics on interpretable models.

The 7PCL aims to detect malignant lesions (melanoma or other). We define a binary classification task on the Atlas test set: predicting whether a lesion is malignant or benign. In Table 3, we compare our baseline model with the interpretable version. None of them has been explicitly trained on this task: the baseline predicts 10 different classes, while the interpretable model predicts 7 signs on which the 7PCL decision rule is applied. Interestingly, training the interpretable model under MSE+CE is beneficial in terms of diagnosis vs using CE only, even if individual sign prediction is lower. As expected [13], the interpretable model suffers from a drop in performance compared to the baseline.

Model	Baseline	Interpretable	CE-only
2-class acc.	85.2%	76.8%	72.8%

Table 3: Models comparison on binary classification.

4 Conclusion

We presented some developments to explain the predictions of our model for skin lesion classification. First, we aim to ensure the model uses relevant medical concepts, not the dataset's spurious correlations. Second, since this tool is developed for practitioners, we seek to enhance their trust in AI. Therefore, we based our effort on two rules used by dermatologists: the ABCD rule and the 7-point checklist.

We derived criteria from the ABCD rule, both visual and numerical, using non-neural algorithms to provide additional information for the neural net classification. We showed that our model uses the medical concepts in the 7PCL, and we derived a more interpretable model predicting each of the seven signs. As expected, a trade-off appeared between explainability and performance in the binary classification of malignant lesions from benign ones.

Future work

This paper presents our first efforts toward a more interpretable skin cancer detection model, which can be extended in several directions. First, doctors could be interviewed about the usefulness of providing additional information to the class prediction with our ABCD-derived visual criteria. Second, some of the performance loss of the interpretable model may be recovered without losing explanations by fitting a residual term [23] based on non-7PCL information. Moreover, this model is limited to binary classification, so it is unable to distinguish skin lesion classes. Extending it will require more data annotated with concepts. Some masks are available in the ISIC dataset but are not as precise as the 7PCL. This could be handled with missing information training.

Acknowledgments

This project is supported by Belle.ai and Torus.ai and European Union's Horizon 2020 research and innovation program through Intelligent Total Body Scanner for Early Detection of Melanoma (iToBoS, grant agreement No. 965221).

References

[1] G Argenziano, H Soyer, V De Giorgi, and et al. Dermoscopy: a tutorial. edra, 2002.

[2] Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, and et al. 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.

[3] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[4] Titus J Brinker, Achim Hekler, Axel Hauschild, and et al. Comparing artificial intelligence algorithms to 157 german dermatologists: the melanoma classification benchmark. *European Journal of Cancer*, 111:30–37, 2019.

[5] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *MethodsX*, 7:100864, 2020.

[6] David Gutman, Noel CF Codella, Emre Celebi, and et al. 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.

[7] Qishen Ha, Bo Liu, and Fuxu Liu. Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge. *arXiv preprint arXiv:2010.05351*, 2020.

[8] Holger A Haenssle, Christine Fink, Roland Schneiderbauer, and et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, 29(8):1836–1842, 2018.

[9] Katja Hauser, Alexander Kurz, Sarah Haggenmüller, and et al. Explainable artificial intelligence in skin cancer recognition: A systematic review. *European Journal of Cancer*, 167:54–69, 2022.

[10] Mohamed A Kassem, Khalid M Hosny, Robertas Damaševičius, and Mohamed Meselhy Eltoukhy. Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review. *Diagnostics*, 11(8):1390, 2021.

[11] 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, 2018.

[12] Been Kim, Martin Wattenberg, Justin Gilmer, and et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[13] 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.

[14] Erik Meijering, M Jacob, J-CF Sarria, and et al. Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 58(2):167–176, 2004.

[15] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, and et al. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.

[16] Darrell S Rigel and John A Carucci. Malignant melanoma: prevention, early detection, and treatment in the 21st century. *CA: a cancer journal for clinicians*, 50(4):215–236, 2000.

[17] Cliff Rosendahl, Alan Cameron, Ian McColl, and David Wilkinson. Dermatoscopy in routine practice: Chaos and clues. *Australian Family Physician*, 41(7):482–487, 2012.

[18] Mohammad Shorfuzzaman. An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. *Multimedia Systems*, 28(4):1309–1323, 2022.

[19] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, and et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[21] Philipp Tschandl, Giuseppe Argenziano, Majid Razmara, and Jordan Yap. Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features. *British Journal of Dermatology*, 181(1):155–165, 2019.

[22] Yiqi Yan, Jeremy Kawahara, and Ghassan Hamarneh. Melanoma recognition via visual attention. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, Proceedings 26*, pages 793–804. Springer, 2019.

[23] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2022.