Towards Domain-Specific Explainable AI: Model Interpretation of a Skin Image Classifier using a Human Approach

Fabian Stieler
University of Augsburg
fabian.stieler@uni-a.de

Fabian Rabe
University of Augsburg

Bernhard Bauer
University of Augsburg

Abstract

Machine Learning models have started to outperform medical experts in some classification tasks. Meanwhile, the question of how these classifiers produce certain results is attracting increasing research attention. Current interpretation methods provide a good starting point in investigating such questions, but they still massively lack the relation to the problem domain. In this work, we present how explanations of an AI system for skin image analysis can be made more domain-specific. We apply the synthesis of Local Interpretable Model-agnostic Explanations (LIME) with the ABCD-rule, a diagnostic approach of dermatologists, and present the results using a Deep Neural Network (DNN) based skin image classifier.

1. Introduction

Skin cancer detection is a popular application for clinical decision support [7]. Motivated by the increasing number of skin cancer patients and the promising therapeutic results for early detection, a lot of research has been done in this field over the past few years. In this context, DNNs have been established as a viable method in the task of developing a model for classifying skin images [2, 8, 12, 30].

The high attention in the community has led to a variety of different approaches with varying levels of performances.[1] Common to all is training a model that can be used for diagnosis and thus for clinical decision support. Consequently, the new approaches have often been evaluated in terms of whether they enable models that lead to better performance results in various dermatological diagnostic tasks [21]. At the same time interpretation of model predictions is increasingly being considered in other areas of AI research. In contrast, the application of these techniques in a skin image classification setting has hardly been addressed, although some recent work has recognized the need [5, 11, 30]. DNNs are known to be opaque and are therefore considered black box models. For their use in critical environments such as the medical field, methods from Explainable AI are needed. Here, like the model itself, the explanations must be adapted to the problem in order to be useful for the particular use case [17].

In this paper, we present a domain-specific idea for this purpose. Our approach synthesizes the machine learning model interpretation methodology LIME [22] with the ABCD rule of dermatoscopy [26], a human diagnosis procedure for distinguishing melanocytic and non-melanocytic skin lesions. We modify the perturbation algorithm of LIME along the dimensions of the ABCD rule and hypothesize predictions of the black box model as presented in section 3. In addition to medically relevant dimensions, medically irrelevant perturbations are introduced to validate the degree of importance of the explanation. Observations are shown in section 4 and discussed in section 5 on a selection of test images from the HAM10000 data set [28]. But first, we provide a brief overview of related work and its methodology.

2. Domain-Specific Explainable AI

Explainable AI (XAI) is a growing field of research that focuses on making a model's decisions understandable. As a result, many innovative techniques recently emerged to help with the interpretation of model behavior [6, 13, 14, 18, 29, 31]. Due to the novelty of this research field, rather generic but hardly problem domain-specific approaches have been developed so far, although the necessity of customized explanations is acknowledged [3, 24].

Studies from a psychological and philosophical perspective has also shown that people are more likely to accept a system if it can explain itself in a way they can understand [17]. The objective is to develop AI systems that can explain decisions in the same way humans do. Linking a machine-aided technique with a human explanation approach can aid us to achieve this goal.

[1] An overview is available by the published results from the ISIC Challenges: https://challenge.isic-archive.com

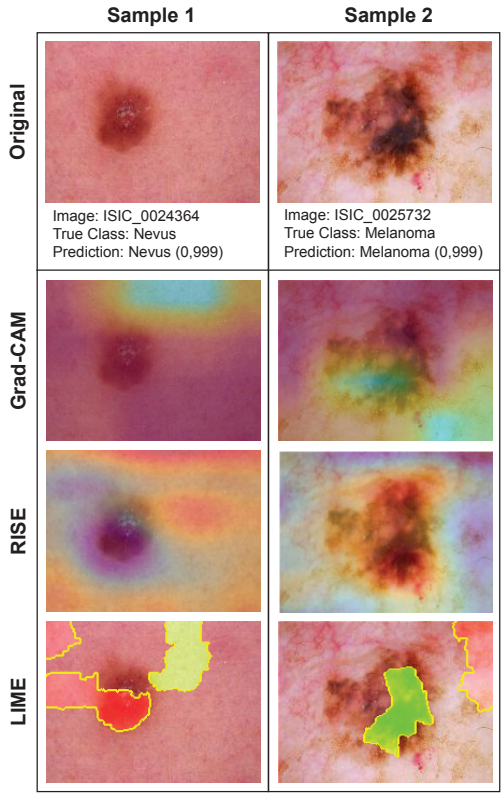

Figure 1. Local Explanations of a DNN-based classifier. Correct model predictions of two test samples are explained by three different model interpretation methods. Colored overlays indicate the degree of importance in relation to the predicted class.

2.1. Model Interpretation Methods

Techniques in the field of XAI can be categorized into two high-level approaches: *Ante-hoc* and *Post-hoc*. The former describes methods which are intrinsically interpretable. Initially, an interpretable, inherently transparent model is defined and then trained. Methods for interpretation of predictions from an previously trained black box model are summarized as *Post-hoc* explanations. Driven by the widespread use of DNN-based skin image classifiers, we start to study methods that support this model type. Furthermore, we establish that a domain-specific approach should allow local explanations for individual predictions.

Gradient-weighted Class Activation Mapping (Grad-CAM) [25] is one of the suitable interpretation methods for this type of model and explanation. However, this function-based approach is limited to Convolution Neural Networks and also needs insights into model parameters. The gradient flowing into the last convolution layer is used to highlight regions in the image that are important for its prediction.

Randomized Input Sampling for Explanation (RISE) [20] is a model-agnostic approach to generate local explanations for image data, based on the principle of occlusion. First, random masks are generated to cover the image areas (pixels) for a given sample. To create an explanation, the sample is occluded with these masks and acquire model predictions. Results are combined by computing the importance of each pixel of the input image with respect to the resulting classification.

An equally common technique are surrogate models, as used in LIME [22]. Here, a data set with perturbed instances is generated for the sample to be explained. The received predictions of the perturbed data using the black box model are weighted and an interpretable local model is trained. For image data this will also work by occluding areas of a given sample. By default, the choice of such sub-regions is performed with super-pixels using the LASSO algorithm, which can lead to the generation of potentially useless subsections, especially in a medical context [16].

This challenge can also be found in Xiang and Wang's research [30], which focuses on interpretable skin image analysis. They introduce an additional stage in a deep learning model training pipeline, and apply LIME to a skin image classification model. It is clarified that such a model interpretation method is able to show meaningful areas in a given sample, but it may lack in specificity for both machines and humans.

In figure 1, three introduced methods were applied to a DNN model. Although the classification of the model for the two shown test samples is highly likely to be correct, the outputs of the explanations reveal a fatal correlation: In case of Sample 2, relevant areas of melanoma are marked. Sample 1 shows a nevus and areas important to the model are outside the lesion. A crucial feature seems to be the skin, the nevus does not contribute to the prediction.

A plain application of such model interpretation methods in an AI system already shows their potential. Instead of relying on the raw prediction, the outputs disclose how the underlying model has come to its decision. However, all these outputs of the different methods show only image areas whose informative value varies considerably. There is a clear lack of domain-specific contexts: To interpret the outcome of these methods, a significant educational effort for domain experts is required.

2.2. Dermatologist's Human Approach

There are a variety of methods for melanoma detection by human pattern recognition. One of the first easy-to-understand frameworks for medical examination and self-examination was introduced in 1985 [10]. Developing this method further, it was later published as the nowadays commonly known ABCD rule of dermatoscopy by Stolz *et al.* [26], which was evaluated 1994 by Nachbar *et al.* [19] and 2010 by Rigel *et al.* [23]. Achieved performance values are reported in these papers with a sensitivity of $\approx 84\%$ and a

specificity of $\approx 83.5\%$.

A comparison of different human approaches compared to a selection of machine-augmented pattern recognition is given by Garau *et al.* [11]. They have illustrated that the ABCD rule outperforms most of the other human- as well as machine learning approaches on the receiver operator characteristic (ROC) curves.

The medical algorithm for visually distinguishing melanocytic and nonmelanocytic lesions is based on multivariate analysis of four criteria. A score is calculated using the properties asymmetry Ⓐ, abrupt truncation of the pigment pattern at the border Ⓑ, different colors Ⓒ and different structural components Ⓓ [19]. In simplified terms, the lesion is examined for all four criteria separately. The higher the score of a criterion applies to the lesion, the more likely it is to be classified as melanocytic. The sum of the scores finally leads to a diagnosis.

Thus, the ABCD rule is particularly suited for use as an human-friendly explainability method for two reasons: First, this approach not only leads to accurate classifications, it is easy to understand for humans, which means that it can be applied not only by physicians, but also to a certain extent by patients themselves. Second, the characteristics used to classify the lesion can be scored independently. Conversely, this has the effect that the four ABCD dimensions can be studied independently. In theory, adding or removing features in the dimensions has a direct impact on the classification.

3. Explainer for Skin Image Classifier

We present the fusion of a model interpretation method with the previously introduced human medical algorithm. Ribeiro's approach in LIME[22], which is on the one hand suitable for image data and on the other hand model-agnostic, tempts to follow the perturbation-based strategy.

An explanation generated by LIME is the minimization of a function considering the complexity Ω of the interpretable model g. $\Omega(g)$ should be as low as possible to be interpreted by a human and is largely determined by the number of features K. Our domain-specific explainer combines the two methods replacing LIME's standard perturbation logic with the criteria of the ABCD rule. Instead of selecting image areas with K super-pixels and then occluding them, we modifying skin images along K diagnostic characteristics.

3.1. Perturbation Dimensions

Scoring each characteristic separately leads to manipulating only one respective dimension in the input image and not changing any features regarding to another dimension. To ensure this, we start with two of the four dermatoscopic-dimensions for our explainer and define them as its medically relevant features: Ⓑ Boundary and Ⓒ Color.

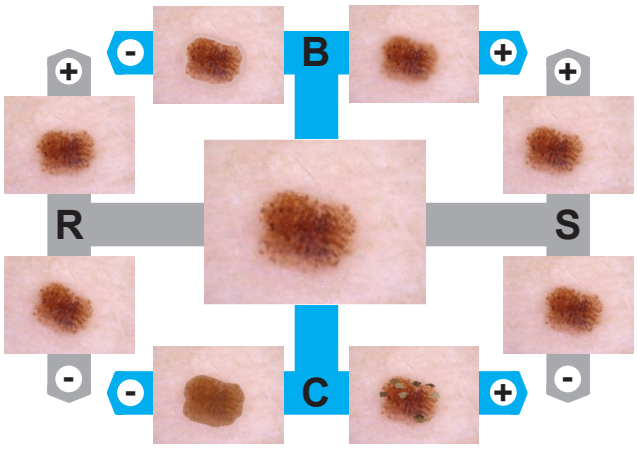

Figure 2. Perturbation dimensions of the explainer. The original image in the center is perturbed along medically relevant (blue) dimensions Ⓑ Boundary and Ⓒ Color, as well as medically irrelevant (gray) dimensions Ⓡ Rotate and Ⓢ Shift, each in a reinforcing (**positive**) and weakening (**negative**) manner.

Following Fong and Vedaldi's research [9] we add two further dimensions to investigate the degree of importance of the explanation. For this the original image is perturbed in a medically irrelevant way without touching any medically relevant features: Ⓡ Rotate and Ⓢ Shift.

Figure 2 illustrates four dimensions with perturbed images in their strongest manifestations of each dimension.[2] These manipulated images are artifacts and may look artificial to a human. However, we have to recognize that the particular characteristic is to be exaggerated. In the following, we go into detail how the perturbation is generated.

Ⓑ Boundary The implementation of the medically relevant dimension is realized along the **negative Boundary** direction by extracting the border area of the segmentation and drawing a sharply delineated line around the lesion. The color of this line corresponds to the average color values of the surrounding image areas and it is ensured that no artifacts arise in relation to the color which is used.

To influence in the **positive Boundary** direction, the edge region is extracted from the segment and a Gaussian blur is added. This causes pixel values to fade into each other and the transition between lesion and skin is less sharply delimited.

Ⓒ Color In the perturbed images of the **negative Color** dimension, the area within the segmentation of the lesion is turned into a uniform color. Possible color irregularities are thus harmonized. The coloring is transparent such that possible structures in the lesion are kept intact.

[2] All image manipulations were implemented with scikit-image: https://scikit-image.org

Adding random color patches in the lesion area produces variation for the **positive Color** direction. They vary in size and color, while ensuring that the color patches are transparent and possible structures remain recognizable similar to the procedure towards negative direction. The chosen colors correspond to plausible shades of brown.

Ⓡ Rotate This perturbation dimension is realized by rotating the sample by a given range of degrees. The range of values corresponds to the **positive** (left) and **negative** (right) direction. We chose mode *'reflect'* as padding strategy, which mirrors neighboring pixel values along the vector.

Ⓢ Shift An affine transformation is performed to shift the skin image. The translation parameter indicates the direction, which is increased with strengthening in the **positive** (left) or **negative** (right) direction. Same as for rotation, *'reflect'* is used to pad the resulting gaps.

3.2. Hypotheses

To further simplify the problem space, we limit the classes studied to nevus *(nv)* and melanoma *(mel)*. The former describes benign neoplasms of melanocytes. In contrast to melanoma, they are usually symmetrical in terms of distribution of color and structure. Melanomas, on the other hand, are defined as malignant neoplasms, which can occur in different variants. [28]

There are medically relevant features, where positive perturbation d^+ means transforming the sample towards melanoma and negative perturbation d^- means to reduce possible features of melanoma in the image. Medically irrelevant perturbations d' neither take away nor add important characteristics. Let y be the probability value of which class a given input sample x belongs to, the following hypotheses about the black box model f received predictions $\hat{y} = \frac{1}{n}\sum_{i=1}^{n} f(x')$, with n perturbed inputs x' by relation $\sim_d$, can be derived:

$A_1^{(nv)}$ Prediction for *nevus* will **decrease** with **positive** perturbation:

$$A_1^{(nv)}(x, x', f) = \{x \sim_{d^+} x' \Rightarrow y > \hat{y}\}$$

$A_0^{(nv)}$ Prediction for *nevus* will increase or remain unchanged with positive perturbation.

$B_1^{(nv)}$ Prediction for *nevus* will **increase** with **negative** perturbation:

$$B_1^{(nv)}(x, x', f) = \{x \sim_{d^-} x' \Rightarrow y < \hat{y}\}$$

$B_0^{(nv)}$ Prediction for *nevus* will decrease or remain unchanged with negative perturbation.

The hypotheses are valid regardless of which dimension of the medically relevant dimensions the perturbations belong to. However, they depend on the given input sample and are therefore not independent of the true class of the sample. Since we are analyzing a two-class problem, the hypotheses for melanoma $[A_1^{(mel)}, A_0^{(mel)}, B_1^{(mel)}, B_0^{(mel)}]$ hold in reverse formulation. In other words, we assume that the negative perturbation of the sample should move the prediction to nevus, while the positive perturbation moves the prediction to melanoma.

Medically irrelevant dimensions should be independent of both the true class of the original image and the dimension to which the perturbations belong. We therefore hypothesize the following:

C_1 The black box model is **inherent** to medically irrelevant perturbations:

$$C_1(x, x', f) = \{x \sim_{d'} x' \Rightarrow f(x) = f(x')\}$$

C_0 Perturbation along medically irrelevant dimensions have significant effects on predictions.

3.3. Experimental Setup

Previously presented hypotheses will be tested with a DNN-based skin image classifier. Therefore, a model was trained with the HAM10000 data set [28], a collection of multi-source dermatoscopic images of common pigmented skin lesions. It was used in the ISIC Skin Lesion Classification Challenge for the past years as well as in numerous studies to train a DNN. In addition to the dermatoscopic images, our explainer takes associated segmentation data [27] as input such that the perturbation can be limited to the lesion.

As already successfully established in other studies [2, 8, 12, 30], we use the transfer learning approach and train a pre-trained MobileNet model [15] with skin image data, which is subsequently able to distinguish between the two relevant classes. Tschandl's data set includes, among other classes, 6,705 images of nevi and 1,113 samples of melanoma. We agree on the annotations assigned by dermatologists as ground truth and split them into training and test data in an 80/20 ratio.

The model performances achieved on the test data can be found in table 1. Obviously, the performance can be im-

	Nevus	Melanoma	Total
Number of Samples	$1,354$	216	$1,561$
True Positives	$1,150$	144	$1,294$
False Positives	203	72	275
F_1-Score	≈ 0.91	≈ 0.57	≈ 0.74*

Table 1. Evaluation results of the classifier. To ensure that class imbalances have no influence, * 'macro' is specified as F_1 average.

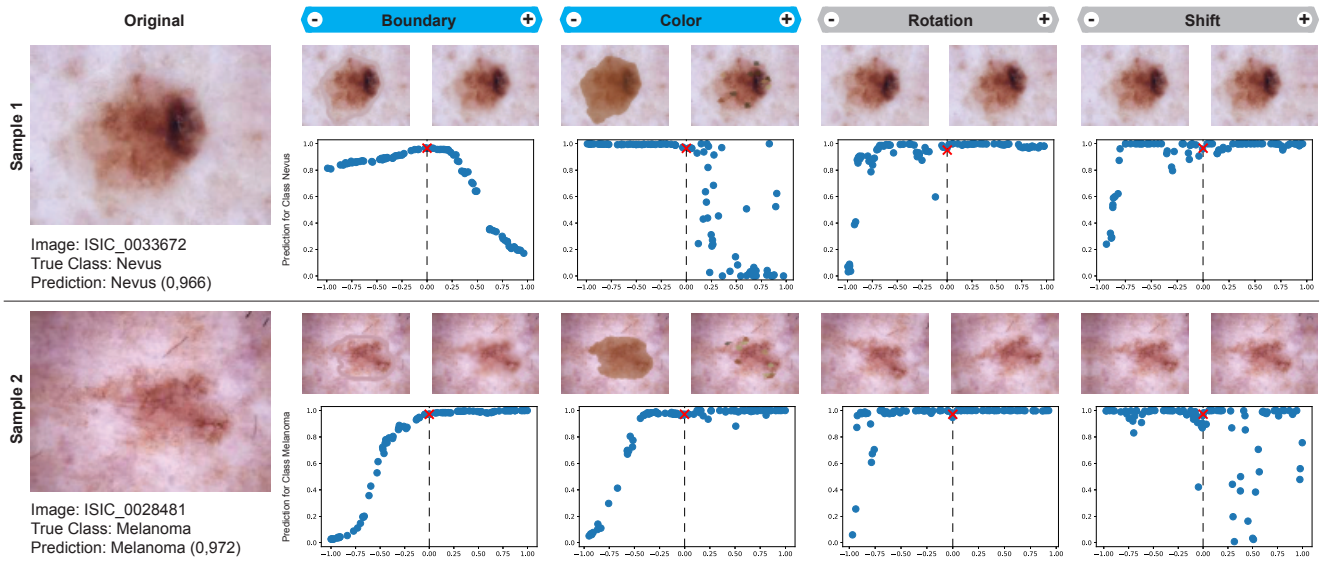

Figure 3. Two original samples, both **correctly** classified **(True Positives)**, with their maximum perturbations for all four explanation-dimensions. Scatter plots under the perturbed images show the prediction of the black box model, each acquired along the indicated dimension.

proved, however, we have deliberately avoided feature engineering and all model tuning techniques for this study in order not to influence the raw output of our explainer in any way.

4. Empirical Results

To study the presented domain-specific explainer in more detail, explanations were generated on a selection from the test images. In order to discuss the results, selected samples with high confidence in the true positive case and low confidence in the false negative case are shown in figures 3 and 4. Another selection criterion was that a significant class flip manifests in at least one dimension.

Both figures can be read according to the following scheme: Respective dimension values are indicated as a heading and the respective maximum perturbed images are shown below them. For each of the samples, scatter plots can also be found in each dimension. Along the ordinate, the prediction value of the black box model is related to the strength of the perturbation, which is plotted along the abscissa.

The scale of the prediction in all scatter plots is set to $[0; 1]$ and refers to the respective class to which the sample corresponds. The strength of the perturbation follows a scale of $[-1; 1]$, which indicates values in the negative value range correspond to the negative perturbation dimension and correspondingly in the positive value range to the positive perturbation dimension.

Furthermore, the scatter plots are separated by a dashed vertical line at position 0. The y-value, represented by a red cross, reflects the classifier's prediction $f(x)$ for the non-perturbed original image. For each of the input samples, $n = 50$ perturbed samples x' in negative as well as positive direction, were generated.

4.1. True Positives

The first case examines the model explanations for correctly classified samples and with our domain-specific approach we try to answer the question "In which dimensions does the model remain accurate?" by using two samples in figure 3 to test the hypotheses.

As observed for **Sample 1**, the prediction in the medically relevant dimensions *Boundary* and *Color* decreases in the area of positive perturbation, which is why we accept $A_1^{(nv)}$ and reject $A_0^{(nv)}$. Regarding the other perturbation direction, $B_1^{(nv)}$ can only be accepted for the *Color* dimension, since in the case of negative perturbation the prediction has a constant value. In the *Boundary* dimension, on the other hand, $B_1^{(nv)}$ must be rejected and we accept $B_0^{(nv)}$ since the prediction does not stagnate or increase but decreases at a constant rate.

To test the third hypothesis C, we look at the two medically irrelevant dimensions. Although the prediction changes at individual perturbation points, we can still observe that it remains at a high level in both the positive and negative value ranges. The average prediction value over all perturbation values is $\hat{y} = 0.931(-0.035)$ for *Rotation* and $\hat{y} = 0.959(-0.007)$ for *Shift*. Values in parentheses denote

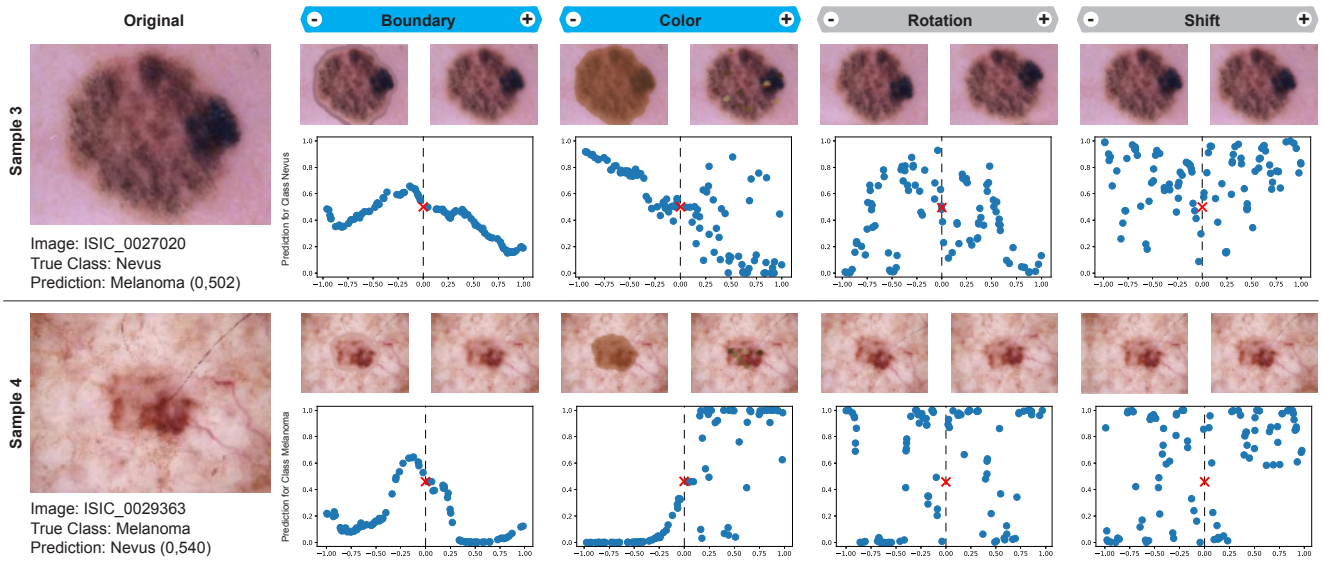

Figure 4. Two original samples, both **incorrectly** classified **(False Positives)**, with their maximum perturbations for all four explanation-dimensions. Scatter plots under the perturbed images show the prediction of the black box model, each acquired along the indicated dimension.

the differences from the prediction of the non-perturbed image and due to the low deviation we decide to accept hypothesis C_1 and reject C_0.

For **Sample 2**, we need to reverse the statements in the hypotheses, since the sample is a melanoma. Now, the prediction for perturbation in the positive direction should increase or remain constant, while the prediction for the negative direction should decrease. These observations are manifested with both medically relevant dimensions *Boundary* and *Color*, which is why we accept both $A_1^{(mel)}$ as well as $B_1^{(mel)}$.

Similar to **Sample 1**, in the medically irrelevant dimensions we can observe that the prediction of the classifier changes along the perturbations. We therefore recalculate the average prediction along all perturbation variables as $\hat{y} = 0.971(-0.001)$ for *Rotation*, and $\hat{y} = 0.907(0.065)$ for *Shift*. The deviation in the prediction allows us to accept C_1 for *Rotation*, but hypothesis C_1 for *Shift* is not supported, so we accept C_0.

4.2. False Positives

The second case investigates model explanations for incorrectly classified samples. Two of such test images are shown in figure 4. An explanation in this scenario is intended to help answer the question, "Why did the model fail?".

As can be clearly seen in the positive direction for **Sample 3** in *Boundary*, we can accept $A_1^{(nv)}$, however $B_1^{(nv)}$ is rejected with regard to the negative direction. $B_1^{(nv)}$, on the other hand, is accepted in the *Color* dimension. The prediction shows both higher and lower values in the positive direction, thus violating the formulation of $A_1^{(nv)}$. However, we decide to accept this hypothesis as well, since the average prediction in the positive direction $\hat{y} = 0.318(-0.184)$ is significantly lower than the prediction of the non-perturbed input image. When testing hypothesis C, it is noticeable at first glance at the *Rotation* and *Shift* dimensions that C_1 must be rejected. The prediction of the classifier shows very different values along all perturbation variables.

With respect to the hypothesis tests of **Sample 4**, it can be seen that for *Boundary* both $A_1^{(mel)}$ and $B_1^{(mel)}$ have to be rejected. Perturbation in both positive and negative direction results in a decreasing prediction. The situation is different for the distribution of the *Color* dimension. Hypothesis $A_1^{(mel)}$ can be accepted at first view because of the decreasing prediction values in the negative area. However, the evaluation of the positive area is less clear, since the values fluctuate along the ordinate here as well. As the average prediction value $\hat{y} = 0.688(+0.148)$ is significantly higher than the prediction of the non-perturbed sample, we decide to accept hypothesis $B_1^{(mel)}$ and to reject $B_0^{(mel)}$.

Looking at the medically irrelevant dimensions, similar observations can be made for **Sample 3**, which is why C_1 is also rejected. It is apparent that the prediction behaviors do not seem to follow a clear pattern, which may be related to the weak model performances for both samples. Yet, the output of the explainer is still helpful as it provides insights

into the model not being robust here and this can help improve both the decision making and the re-development of the model.

5. Discussion

Empirical results presented above demonstrate how the black box model responds to the perturbed images and allows us to draw conclusions about which features may have been important. However, this information still needs to be translated into an *explanation* for the user.

The model behavior was observed with only one trained model. Results may differ with various model architectures and training data sets. Additionally, it should be noted that the perturbation of input samples in the results is successive and simultaneous perturbation of multiple dimensions remains to be investigated.

Our goal was to find a domain-specific approach for local explanations of a DNN model. Besides the explanation for a single sample, global model explanations provide indispensable insights. A first step in this direction could be the application of our explanation method to multiple instances and the subsequent aggregation of the results.

In addition, there is missing evidence between the observed importance of feature-dimension and the true score according to the ABCD rule. This, along with measurement of other metrics to evaluate explanations, leaves room for future research.

6. Conclusion and Future Work

The skin image classifier in a clinical decision support system can serve as a second opinion for a dermatologist. To a limited degree, the strong research community has already made it possible to realize such tasks today. However, the models in such an AI-based system for the dermatologist only provide predictions, but the physician cannot ask why the classifier came up with its decision.

XAI methods are intended to meet this need. We have shown how a domain-specific approach for skin image analysis can look like. A conceivable scenario would be that a dermatologist diagnoses a lesion as a nevus, but the model classifies it as a melanoma. This mismatch leaves the treating physician (and his patient) in a skeptical position, which is why both people ask: "Why?" With our approach, the answer could be: "If the color is harmonized in the lesion, the confidence with respect to the given prediction of the classifier decreases". The physician recognizes a color irregularity in the dermatoscopic image, which is not visible on the lesion, and can therefore explain why the classifier incorrectly tended to diagnose melanoma.

The physician is either confirmed in its diagnosis by a clinical decision support system or a contrary diagnosis is made by the system. In both cases, it is enormously helpful if human-understandable explanations can be generated to interpret the predictions. Approaches adapted to the respective domain not only create more trust, but also a greater understanding of the system.

Following the current results from sections 4 and 5, future work may study the two remaining medically relevant dimensions, asymmetry and differential structure, for which the work by Barata *et al*. provides a overview of feature extraction in dermoscopy image analysis [4]. Ali *et al*. show a way to extract these features from lesions [1]. Closely resembling what Almaraz-Damian *et al*. have shown in their paper [2], another possible task could be to use the data perturbed by our explainer as training data to investigate both the performance of the resulting model and whether the predictions follow a different pattern.

Moreover, the approach of perturbation-based explanations using medically relevant and medically irrelevant features for diagnosis may be applicable in other medical specialties.

7. Acknowledgement

This work was funded by the German Federal Ministry of Education and Research (BMBF) under reference number 031L9196B.

References

[1] Abder-Rahman Ali, Jingpeng Li, and Sally Jane O'Shea. Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images. *PLOS ONE*, 15(6):e0234352, 2020. 7

[2] Jose-Agustin Almaraz-Damian, Volodymyr Ponomaryov, Sergiy Sadovnychiy, and Heydy Castillejos-Fernandez. Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. *Entropy*, 22(4):484, 2020. 1, 4, 7

[3] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv:1909.03012 [cs, stat]*, 2019. 1

[4] Catarina Barata, M. Emre Celebi, and Jorge S. Marques. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1096–1109, 2019. 7

[5] Catarina Barata, M. Emre Celebi, and Jorge S. Marques. Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition*, 110:107413, 2021. 1

[6] Vaishak Belle and Ioannis Papantonis. Principles and practice of explainable machine learning. *arXiv:2009.11698 [cs, stat]*, 2020. 1

[7] M. Emre Celebi, Noel Codella, and Allan Halpern. Dermoscopy image analysis: Overview and future directions. *IEEE Journal of Biomedical and Health Informatics*, 23(2):474–478, 2019. 1

[8] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 1, 4

[9] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, 2017. 3

[10] Robert J. Friedman, Darrell S. Rigel, and Alfred W. Kopf. Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians*, 35(3):130–151, 1985. 2

[11] Daniel S. Gareau, James Browning, Joel Correa Da Rosa, Mayte Suarez-Farinas, Samantha Lish, Amanda M. Zong, Benjamin Firester, Charles Vrattos, Yael Renert-Yuval, Mauricio Gamboa, María G. Vallone, Zamira F. Barragán-Estudillo, Alejandra L. Tamez-Peña, Javier Montoya, Miriam A. Jesús-Silva, Cristina Carrera, Josep Malvehy, Susana Puig, Ashfaq Marghoob, John A. Carucci, and James G. Krueger. Deep learning-level melanoma detection by interpretable machine learning and imaging biomarker cues. *Journal of Biomedical Optics*, 25(11), 2020. 1, 3

[12] 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. 1, 4

[13] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *arXiv:1806.00069 [cs, stat]*, 2019. 1

[14] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models. *arXiv:1802.01933 [cs]*, 2018. 1

[15] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861 [cs]*, 2017. 4

[16] Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. *Computers in Biology and Medicine*, 126:104041, 2020. 2

[17] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38, 2019. 1

[18] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. 1

[19] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B. Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. 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. 2, 3

[20] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. *CoRR*, abs/1806.07421, 2018. 2

[21] Michael Phillips, Jack Greenhalgh, Helen Marsden, and Ioulios Palamaras. Detection of malignant melanoma using artificial intelligence: An observational study of diagnostic accuracy. *Dermatology Practical & Conceptual*, page e2020011, 2019. 1

[22] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. 1, 2, 3

[23] Darrell S. Rigel, Julie Russak, and Robert Friedman. The evolution of melanoma diagnosis: 25 years beyond the ABCDs. *CA: A Cancer Journal for Clinicians*, 60(5):301–316, 2010. 2

[24] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020. 1

[25] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 2

[26] W Stolz, D Hölzel, A Riemann, W Abmayr, C Przetak, P Bilek, M Landthaler, and O Braun-Falco. Multivariate analysis of criteria given by dermatoscopy for the recognition of melanocytic lesions. In *Book of Abstracts, Fiftieth Meeting of the American Academy of Dermatology, Dallas, Tex: Dec*, pages 7–12, 1991. 1, 2

[27] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020. 4

[28] 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(1):180161, 2018. 1, 4

[29] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review, 2020. 1

[30] Alec Xiang and Fei Wang. Towards interpretable skin lesion classification with deep learning models. *AMIA Annual Symposium proceeding*, pages 1246–1255, 2019. Publisher: American Medical Informatics Association. 1, 2, 4

[31] Quanshi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *arXiv:1802.00614 [cs]*, 2018. 1