Conditional dependence tests reveal the usage of ABCD rule features and bias variables in automatic skin lesion classification

Christian Reimers[1,2], Niklas Penzel[1], Paul Bodesheim[1], Jakob Runge[2,3], Joachim Denzler[1,2]
[1] Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany
[2] Institute of Data Science, German Aerospace Center, Jena, Germany
[3] Technische Universität Berlin, Berlin, Germany
{christian.reimers, niklas.penzel, paul.bodesheim, joachim.denzler}@uni-jena.de,
jakob.runge@dlr.de

Abstract

Skin cancer is the most common form of cancer, and melanoma is the leading cause of cancer related deaths. To improve the chances of survival, early detection of melanoma is crucial. Automated systems for classifying skin lesions can assist with initial analysis. However, if we expect people to entrust their well-being to an automatic classification algorithm, it is important to ensure that the algorithm makes medically sound decisions. We investigate this question by testing whether two state-of-the-art models use the features defined in the dermoscopic ABCD rule or whether they rely on biases. We use a method that frames supervised learning as a structural causal model, thus reducing the question whether a feature is used to a conditional dependence test. We show that this conditional dependence method yields meaningful results on data from the ISIC archive. Furthermore, we find that the selected models incorporate asymmetry, border and dermoscopic structures in their decisions but not color. Finally, we show that the same classifiers also use bias features such as the patient's age, skin color or the existence of colorful patches.

1. Introduction

Skin cancer and especially malignant melanoma is a dangerous and common form of cancer. Diagnosis in the early stages is essential to improve the survival rate [13]. To ensure an early discovery of cancer, patients need to undergo regular checks by trained medical professionals. However, there is not enough medical personnel in most regions to comprehensively offer such a labor-intensive examination. One possibility to reduce the amount of human labor needed is the employment of automatic skin lesion classifiers.

Different approaches for automatic skin lesion classification exist to support practitioners [29, 14, 12, 5, 6]. One approach is to automatically extract handcrafted features that are deemed important by dermatologists, as proposed, for example, by [19, 7]. These works concentrate on the features named in the dermoscopic ABCD rule [42, 27] to earn the trust of dermatologists and patients.

The ABCD rule is an algorithm for dermatologists to differentiate between melanoma and nevi in dermoscopic skin lesion images. To this end, the dermatologist scores four features that give the rule its name. The four features are: **A**symmetry, **B**order, **C**olor and **D**ermoscopic structures. Section 4.3 describes how these features are scored. After deriving individual feature scores, dermatologists combine them into a total dermoscopy score. A simple threshold of this total score yields high accuracy to distinguish melanoma from benign nevi.

The algorithmic nature of the ABCD rule allows for straightforward automatization, where the most challenging task is the automatic feature extraction. The resulting skin lesion classifiers are explainable and can be trusted by practitioners and patients alike. However, in recent years, the state-of-the-art in automatic skin lesion classification has shifted away from implementations of the ABCD rule towards large ensembles of very deep neural networks. These networks are often pre-trained on unrelated image datasets and employ heavy test time augmentations, *e.g.*, [14]. On the one hand, this shift allowed researchers to construct automatic skin lesion classifiers that outperform even experienced practitioners [45]. On the other hand, an integral part of deep learning is automatic feature selection [33], meaning that the researcher has no control over which feature the model selects. In particular, it is not straightforward to determine whether an automatic classifier still uses the ABCD rule features or if they heavily rely on bias features, *e.g.*, [24, 36, 26, 4].

To determine which features are relevant to a deep neural network's decision, most researchers have employed

saliency maps, *e.g.*, [47]. Saliency maps are a method that highlights areas of the input image that are relevant to a single decision. However, we discuss in Section 3, why we think that saliency maps are not the best solution in our situation. The most important reason is that features such as asymmetry do not correspond to an image region. Hence, saliency maps can not highlight them. Instead, we recommend to use the method proposed by [35]. This method uses the framework of causality and structural causal models [28]. It has the advantage that it can determine the relevance of features not represented by regions of the input. Therefore, it can determine the relevance of, for example, asymmetry. We provide a detailed description of the method in Section 3.1.

In this paper, we present three results. First, we conduct sanity checks to determine that the method is suitable, as we are the first to apply it to this kind of data. To this end, we investigate four features that contain little to no information relevant for skin lesion classification and demonstrate that the classifiers do not, or very rarely, base their prediction on the information in these features. Second, we show that state-of-the-art classifiers use the asymmetry and the border features defined in the ABCD rule to classify melanoma. They rely on dermoscopic structures when determining whether a skin lesion is seborrheic keratosis. By contrast, the models we analyzed do not use the color feature that dermatologists deem relevant for melanoma classification. Third, we find that the classifiers also use the patients' age and skin color to classify a skin lesion. Both can be estimated from an image, for example, from body hair. Further, we demonstrate that classifiers pick up the spurious connection between colorful patches and nevi in the images of the SONIC dataset [38], a bias also reported by [24, 36].

2. Related work

This work aims to achieve three goals. First, we validate that the method described in [34, 35] can be used on the complex real-life task of skin lesion classification. Second, we determine whether state-of-the-art classifiers use the features listed in the ABCD rule. Third, we investigate whether automatic classifiers take shortcuts by relying on spurious correlations in the data.

To the best of our knowledge, previous work only validates the method of [35] in small- [35], and toy examples [35, 34]. In contrast, we systematically investigate, whether the method falsely indicates the use of features that have little to no information in a real-world scenario.

As far as we know, we are also the first to determine whether state-of-the-art classifiers use ABCD rule features. The closest works from the literature do a more general investigation into which features a classifier relies on. The authors of these works often use saliency maps, which can not determine the relevance of some features such as asymmetry, *e.g.*, [47]. Other related works automatically extract the features named in the ABCD rule and build a classifier upon them [19, 7]. The worse performance of these systems indicates that deep models learn more or other features. By contrast, we evaluate directly if deep classifiers use the features from the ABCD rule.

More related work exists for the critical task of determining whether classifiers are biased. To investigate the influence of spurious biases on their results, the authors of [47] use different saliency methods as local explanations of InceptionNet models [43]. They employ GradCAM [39] as well as Kernel SHAP [23] and observe that models sometimes base their decisions on background information rather than the skin lesion. The overview [24] includes a list of several known artifacts in dermoscopy images. One prominent example is the occurrence of large colorful patches next to skin lesions. These patches, introduced by the SONIC dataset [38], form a bias in the ISIC archive [1]. Using saliency maps, the authors of [36] find that their model for skin lesion classification "looks" at those patches, which may affect the predictions. Both these works use saliency map techniques which are only suitable for certain biases. They can not detect biases such as the patient's age or sex. Moreover, saliency maps leave researchers with the semantic task of interpreting them. The authors of [4] analyze biases introduced into the ISIC archive by removing medical information. They find that models correctly classify images even if there is almost no clinically meaningful information left in the remaining inputs. Thus, they believe that the classifier relies on spurious biases to inflate its performance. In [26], Muckatiera applies a pruning algorithm to a skin lesion classifier and derives multiple subnetworks. The accuracies achieved by these networks differ for varying age-groups and male and female patients. In contrast, the approach presented in this paper offers the possibility to investigate the usage of such features directly. The output of the method [35] needs neither semantic interpretation nor comparison.

3. Method

This section is split into two parts. We first describe the conditional dependence method of [35] and why we chose it over saliency map methods. Since our selected method uses conditional dependence tests, in Section 3.2, we describe the specific tests we employ in our experiments.

3.1. Determining the relevance of features

Deep neural networks rely on automatic feature selection, complicated network architectures, and optimization methods. Their excellent performance leads to the wide adoption of black-box classifiers, which complicates determining which features of the input image influence the decision.

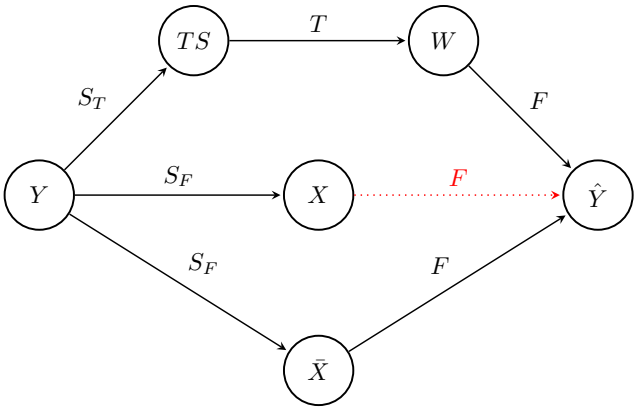

Figure 1: The graph of the structural causal model associated with supervised learning. Ground truth is parametrized by the label Y. From it, we can sample the training set TS using the sampling process S_T. Additionally, we sample the feature X and the class $\bar{X}$ of all features independent of X, using the sampling process S_F. The training process T maps the training set TS onto the set of weights W. These weights and some of the features in $\bar{X}$ are used by the neural network to produce its prediction $\hat{Y}$. The only remaining question is if the neural network uses the feature X. Here Reichenbach's Common Cause Principle [31] can be used. This figure follows Figure 4 in [35].

To this end, we decide to use the method proposed by [35]. The authors frame supervised learning as a structural causal model [28]. They want to determine whether the classifier uses a specific feature. They use this feature of interest together with all orthogonal features as a representation of the input image. Furthermore, they consider the label, the training set, the classifier's weights, and its prediction as variables. The structural causal model additionally incorporates the functions connecting these variables, namely sampling, training, and inference. Figure 1 depicts the resulting graphical model. For further explanations and discussions of this method, we refer the reader to [35].

Using this framing, the question whether a function connects the feature of interest X and the prediction of the classifier $\hat{Y}$ can be answered by testing for the conditional dependence of the two variables

$$X \not\perp\!\!\!\perp \hat{Y} \,|\, Y. \tag{1}$$

If the question is answered affirmative, the only possible function connecting the two variables is the classifier. In this case, we can infer that the classifier uses the feature of interest for its prediction. The feature either speaks for a class or leads, for example, to less clear cut decisions. This framework has certain limitations. First, it can not detect effects that influence an individual while keeping the distribution of all individuals the same. Second, the selection of the independence test is difficult but crucial. Third, the method represents an image as the set of orthogonal features $X \cup \bar{X}$. Semantic features, however, are not orthogonal. Hence, if we test a feature, we have to decide which information we include. The length of the border of a skin lesion, for example, will contain information on its area. In contrast, the border feature we describe in Section 4.3 does not.

Other solutions for the task of determining which features are used by a deep neural network have been proposed. The most popular approach is saliency maps [48, 41, 40, 39, 49, 25, 21]. They assign a relevance value to each pixel. This way, they highlight relevant areas of the input image.

For the task at hand, however, saliency maps are not ideal. A saliency map only explains a single decision of a classifier and does not claim generality. Even further, different authors have demonstrated, that saliency maps lack sensitivity to parameter values [3], that they are unreliable [20], and that they are fragile [15]. The main reason, however, we chose the method of [35] over a saliency map method is that it can be used for features that are not regions of the input. Out of the twelve features we use throughout this work, only the colorful patches can be represented as a region of the input.

3.2. Dependence tests

Since we reduce the question of whether a classifier uses a feature to a conditional dependence test, in the following, we introduce the three dependence tests we are using throughout this work. In addition, we briefly describe their advantages and drawbacks.

Partial correlation (PC) For this test, we need to erase the influence of the labels Y on both X and $\hat{Y}$. Since we condition on a categorical variable, this is simply done by calculating

$$X_{|Y} = X - \mathbb{E}(X \,|\, Y), \qquad \hat{Y}_{|Y} = \hat{Y} - \mathbb{E}(\hat{Y} \,|\, Y). \tag{2}$$

The test statistic is calculated as the coefficient of determination [2] between $\hat{Y}_{|Y}$ and $X_{|Y}$. To check whether this correlation is significant, we perform a shuffle test, meaning that we shuffle all values and calculate the coefficient of determination again. Doing this a thousand times allows us to approximate the distribution of coefficients of determination under the assumption of independence. If our observed value is larger than 99% of these values, we assume that the correlation is significant.

The main disadvantage of this dependence test is that it only captures linear connections between the feature of interest and the prediction. However, the test has two main advantages. First, the test statistic is interpretable. The coefficient of determination states which fraction of the variance in the prediction can be explained by a linear model of the feature of interest. The second advantage is that this

test only detects if the feature speaks for or against a particular class and not, for example, if the feature only leads to a more precise classification. While this is not an advantage in general, we believe that it is in our application.

Fast conditional independence test (FCIT) This test was introduced by [8]. The main idea is to check whether one variable can predict the other variable using a decision tree regressor. The test is fast and can detect non-linear relations. However, it is not a mathematical dependence test, and known failure cases exist [8]. For a more detailed discussion of this test, we refer the reader to [8].

Hilbert-Schmidt independence criterion (HSIC) The HSIC test was proposed by [16] and involves calculating the correlation in a kernel space instead of the input space to account for nonlinear relationships. Similar to the partial correlation, we use $X_{|Y}$ and $\hat{Y}_{|Y}$ as well as a shuffle test to approximate the distribution under the assumption of independence. The test statistic is given by

$$\text{HSIC}(X_{|Y}, \hat{Y}_{|Y}) = \frac{1}{(m-1)^2}\mathbf{tr}\left(K_{X_{|Y}} H K_{\hat{Y}_{|Y}} H\right), \quad (3)$$

with m being the number of examples, $K_{X_{|Y}}$ the kernel matrix associated with a Gaussian kernel for the variable $X_{|Y}$, and $H_{ij} := \mathbf{1}_{i=j} - m^{-2}$ a normalization matrix. Again, we set the level of significance to 0.01. Since the HSIC test requires the calculations of large kernel matrices, we perform the test only on a random subset of 1,000 samples. To counteract this restriction, we use a larger variance for the Gaussian kernel of a ten times the mean distance of all samples. The main advantage of the HSIC test is that it is a true mathematical dependence test that can detect nonlinear dependencies. However, the results depend on the variance of the kernel that has been selected.

4. Experiments

This section discusses the details of our experimental setup. We test whether state-of-the-art skin lesion classification systems use specific features to reach their decision. Since the classifiers are the same in every experiment, we start by describing them in Section 4.1. Our experiments differ in terms of the features we analyze and the datasets we use. We explain the exact features and datasets corresponding to the respective experiments in Sections 4.2, 4.3, and 4.4. The results are summarized in Section 5 and discussed in Section 6.

4.1. Classifiers

In recent years, successful ISIC challenge participants often used deep learning models pre-trained on ImageNet [37], test-time augmentations, ensembles, or combinations thereof [9]. In this paper, we analyze two of these models. We start by describing them in the following.

The first classifier presented by Perez *et al*. [29] introduced test time augmentations and won the best paper award at the ISIC Skin Image Analysis Workshop @ MICCAI 2018. They aggregated the predictions of multiple augmented test examples into one final prediction. We follow their training scheme on the ISIC 2017 challenge dataset [10] and use the augmentation scenario "J" as described in [29]. We train multiple models using different backbones on the binary tasks of melanoma (MEL) and seborrheic keratosis (SK) classification. We report results using mean (n) and maximum (x) aggregation. Further, we use ResNet-152 (R) [17], Inception-v4 (I) [43], and DenseNet-161 (D) [18] as backbone networks. For further details and specific hyperparameter settings, we refer the reader to [29] and the code the authors provided. We abbreviate the different models after the following scheme: Dx26::MEL is a DenseNet-161 trained for melanoma classification using maximum aggregation of 26 augmented examples.

Gessert *et al*. [14] won both tasks of the ISIC 2019 Skin Lesion Classification Challenge. Their classifier is a large ensemble mainly based on multi-resolution EfficientNets [44] pretrained on ImageNet. Similar to Perez *et al*. [29], they also heavily utilize extensive data augmentations. Their classifier performs multi-label classification with eight classes and one rejection class. It is trained on three datasets: HAM10000 [46], BCN20000 [11], and MSK [10]. We train an ensemble of five EfficientNets (B0) following the training scheme of Gessert *et al*. [14]. We refer to [14] and the authors' code for specific details and hyperparameter settings. Following Gessert *et al*. [14], we average the predictions of 36 ordered crops of each example to derive a final prediction. Note that the full model proposed by Gessert *et al*. [14] contains much more and larger networks than just the five EfficientNets (B0). Therefore, we expect our model to perform worse than the original. However, as this is the default parameter setting in the code the authors provided, we expect it to be a decent representation of their classifier.

The classification models of Perez *et al*. [29] and Gessert *et al*. [14] are just examples, but they employ different widely-used strategies also contained in other high-performance models. Hence, we argue that these classifiers are a good representation of current strategies used in skin lesion classification. They also cover the subdivision in melanoma classification and skin lesion classification in general. Performances achieved by these models in our setup can be found in the supplementary material.

4.2. Validation of the conditional dependence method

In the first experiment, we investigate if models use features with no or very little meaningful information. The goal is to validate that the method we chose can handle our

Table 1: Results of the validation of the conditional dependence method. For every classifier and feature/test, we indicate that the feature is used with a ✓or not used with a ✗. We assume a feature is used if the test reports a significant dependence at $p = 0.01$. The models of Perez *et al*. [29] are denoted by their backbone (ResNet-152 (R), Inception-v4 (I), DenseNet-161 (D)) the aggregation method (mean (n), maximum (x)) and the number of augmented samples. For the ensemble of Gessert *et al*. [14] we denote the predicted class as melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC) and squamous cell carcinoma (SCC). The star (*) denotes cases where the labels already explain all of the observed variance.

	Orientation			Rand. Symmetry			Image ID			MNIST Class		
Classification Model	PC	FCIT	HSIC	PC	FCIT	HSIC	PC	FCIT	HSIC	PC	FCIT	HSIC
Perez *et al*. [29]:Dx26::MEL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Perez *et al*. [29]:Dn26::MEL	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:Dn64::MEL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:Ix26::MEL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:In26::MEL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:In64::MEL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:Rx26::MEL	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗
Perez *et al*. [29]:Rn26::MEL	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:Rn64::MEL	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓
Perez *et al*. [29]:Dx26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Perez *et al*. [29]:Dn26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Perez *et al*. [29]:Ix26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗*
Perez *et al*. [29]:In26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:Rx26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:Rn26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Gessert *et al*. [14]::MEL	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗
Gessert *et al*. [14]::NV	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
Gessert *et al*. [14]::BCC	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Gessert *et al*. [14]::AK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Gessert *et al*. [14]::BKL	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Gessert *et al*. [14]::DF	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗*
Gessert *et al*. [14]::VASC	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
Gessert *et al*. [14]::SCC	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

task's complex data and still produce meaningful results. We expect that the classifiers do not use any information contained in the features discussed in this section. If we later detect the usage of features in the other experiments, we know that these detections are meaningful. We do not validate the inverse as there is no consensus on which features have to be used by neural networks. We test the following features:

Orientation: For the orientation feature, we estimate the ellipse with the same second moments as the skin lesion's contour and measure the angle between its major axis and the horizontal axis of the image. Since this feature measures the lesion's orientation in the image, it is a property of the image rather than the lesion. Therefore, it does not contain any useful information for classifying the lesion.

Random symmetry: We first choose a random axis that goes through the center of gravity of the lesion segmentation and calculate the intersection over union (IOU) between the areas of the skin lesion and the skin lesion flipped along this axis. After repeating this process for the orthogonal axis, we multiply both IOU scores to obtain the feature.

Image ID: Here, we take the position of the image in the ISIC archive [1]. Although images from the same source receive consecutive numbers in the archive, this feature contains very little useful information.

MNIST class: We calculate this feature by feeding the segmentation mask of a skin lesion into a classifier for handwritten digits trained on the MNIST dataset [22]. The details of the corresponding classifier are provided in the supplementary material. This feature contains almost no useful information since the similarity between classifying skin lesions and classifying hand-written digits is minimal.

Table 2: Results for the clinically meaningful ABCD rule features. For notation see Table 1.

	Asymmetry			Border			Color			Derm. Structures		
Classification Model	PC	FCIT	HSIC	PC	FCIT	HSIC	PC	FCIT	HSIC	PC	FCIT	HSIC
Perez *et al*. [29]:Dx26::MEL	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
Perez *et al*. [29]:Dn26::MEL	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
Perez *et al*. [29]:Dn64::MEL	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
Perez *et al*. [29]:Ix26::MEL	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✓
Perez *et al*. [29]:In26::MEL	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:In64::MEL	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Perez *et al*. [29]:Rx26::MEL	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✓
Perez *et al*. [29]:Rn26::MEL	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓
Perez *et al*. [29]:Rn64::MEL	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓
Perez *et al*. [29]:Dx26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓
Perez *et al*. [29]:Dn26::SK	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓
Perez *et al*. [29]:Ix26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Perez *et al*. [29]:In26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓
Perez *et al*. [29]:Rx26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Perez *et al*. [29]:Rn26::SK	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓
Gessert *et al*. [14]::MEL	✗	✗	✓	✗	✓	✓	✓	✓	✓	✗	✗	✓
Gessert *et al*. [14]::NV	✗	✗	✓	✓	✗	✓	✓	✓	✓	✗	✗	✓
Gessert *et al*. [14]::BCC	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Gessert *et al*. [14]::AK	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓
Gessert *et al*. [14]::BKL	✗	✗	✓	✓	✗	✓	✗	✗	✓	✓	✗	✓
Gessert *et al*. [14]::DF	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Gessert *et al*. [14]::VASC	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓
Gessert *et al*. [14]::SCC	✗	✗	✓	✓	✗	✓	✗	✓	✓	✗	✗	✓

We choose these features because the first contains no, and the other three minimal useful information on the skin lesion. It is reasonable to assume that they should not or only very rarely be used by the selected classifiers. Furthermore, these meaningless features have a similar structure as the interesting features we want to test later on. For example, orientation and random symmetry have a similar complexity to the asymmetry and border features described in Section 4.3. Additionally, the MNIST class feature's complexity is similar to the dermoscopic structures feature from the same section. We acquire the image ID feature from the metadata similarly to the age and sex features described in Section 4.4.

We evaluate the distributions of these features and the predictions on the HAM10000 dataset [46]. This dataset contains 10,015 images of seven different classes as well as ground truth segmentations and different metadata. Examples include the age and the sex of a patient.

4.3. Clinically meaningful features:

In the second experiment, we investigate the features introduced by the ABCD rule. Dermatologists deem these features helpful to determine whether a skin lesion is a melanoma. Here, we describe how dermatologists score these features and how we automate this scoring.

Asymmetry: This feature's score is the maximum number of orthogonal lines that can be found such that the skin lesion is almost symmetric to all of them. Since there can be at most two orthogonal lines in an image, this score is either zero, one, or two. To automatically evaluate this feature, we use axes that form an integer degree angle with the image's horizontal axis. We consider the lesion to be symmetric if the intersection over union of the lesion area and the area of the lesion flipped along the axis is larger than 0.9.

Border: Dermatologists assess the border by dividing it into eight segments. The number of border segments with a sharp and abrupt separation from the surrounding area defines this feature's score. Hence, the score may take values between zero and eight. Following related work, we instead use the isoperimetric fraction, *i.e*. the fraction of the area A of the skin lesion and the squared length of its perimeter P

$$\text{border} = \frac{4\pi A}{P^2}. \tag{4}$$

Color: Doctors score the color by counting how many of

Table 3: Results for the known bias features. For notation see Table 1.

	Age			Sex			Skin Color			Colorful Patches		
Classification Model	PC	FCIT	HSIC	PC	FCIT	HSIC	PC	FCIT	HSIC	PC	FCIT	HSIC
Perez *et al.* [29]:Dx26::MEL	✓	✗	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓
Perez *et al.* [29]:Dn26::MEL	✓	✓	✓	✗	✗	✓	✗	✓	✗	✓	✓	✓
Perez *et al.* [29]:Dn64::MEL	✓	✗	✓	✗	✗	✗	✗	✓	✗	✓	✓	✓
Perez *et al.* [29]:Ix26::MEL	✓	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	✓
Perez *et al.* [29]:In26::MEL	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓
Perez *et al.* [29]:In64::MEL	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Perez *et al.* [29]:Rx26::MEL	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓
Perez *et al.* [29]:Rn26::MEL	✓	✗	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓
Perez *et al.* [29]:Rn64::MEL	✓	✗	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓
Perez *et al.* [29]:Dx26::SK	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓
Perez *et al.* [29]:Dn26::SK	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓
Perez *et al.* [29]:Ix26::SK	✓	✗	✓	✗	✗	✓	✓	✗	✓	✓	✓	✓
Perez *et al.* [29]:In26::SK	✓	✗	✓	✗	✗	✓	✓	✗	✓	✓	✓	✓
Perez *et al.* [29]:Rx26::SK	✓	✗	✓	✗	✗	✓	✓	✗	✓	✗	✓	✓
Perez *et al.* [29]:Rn26::SK	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓
Gessert *et al.* [14]::MEL	✗	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓
Gessert *et al.* [14]::NV	✓	✗	✓	✗	✗	✓	✓	✓	✓	✗	✓	✓
Gessert *et al.* [14]::BCC	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓
Gessert *et al.* [14]::AK	✓	✗	✓	✗	✗	✓	✗	✓	✓	✗	✓	✓
Gessert *et al.* [14]::BKL	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓
Gessert *et al.* [14]::DF	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓
Gessert *et al.* [14]::VASC	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓
Gessert *et al.* [14]::SCC	✓	✗	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓

the colors white, red, light brown, dark brown, blue-gray, and black appear in the skin lesion. To automatically calculate this score, we defined neighborhoods in the HSV space for each color. The details can be found in the supplementary material. To calculate the color feature, we count for how many of these specified color intervals we can find a pixel of this color in the skin lesion.

Dermoscopic structures: This feature is the number of different structures that appear in the lesion. The possible structures are milia like cysts, negative networks, pigment networks, streaks, and globules. To evaluate this feature, we rely on data labeled for a corresponding task in the 2018 ISIC challenge [9].

The models should incorporate these features in their decisions to increase the trust in automatic classification systems. To test this, we use the HAM10000 dataset for the first three features. To investigate the dermoscopic structures feature, we use the dataset of the 2018 ISIC Challenge [9]. This dataset contains 2,594 images and five segmentation masks for each image corresponding to different dermoscopic structures.

4.4. Known bias features

In the third experiment, we determine if the following bias features influence the predictions.

Age: For this feature, we use the metadata of the HAM10000 dataset. This dataset includes approximations for the patient's age rounded to the nearest five years.

Sex: Similar to age, a patient's sex is also annotated in the HAM10000 dataset. We extract it as a binary variable.

Skin color: To determine the skin color, we use the ten by ten pixel area in the top left corner of the image. We exclude images where this area is black and perform a principal component analysis (PCA) on the extracted area of the remaining images. The loading of the first principal component determines the skin color feature. Visual inspection shows that low values correlate with light skin color. Some examples can be found in the supplementary material.

Colorful patches: Some ISIC archive images, especially images from the SONIC dataset [38], contain colorful patches. The supplementary material contains some example images. Rieger *et al.* [36] note that classifiers might use these patches as features since all images containing at least one colorful patch show benign skin lesions. To investigate

this further, we check whether the patch area as a feature influences the predictions. For this feature, we use the patch segmentations provided by Rieger *et al*. [36].

For the evaluation of the first three bias features, we use the HAM10000 dataset. For the final feature, we instead use the first 10,000 images from the ISIC archive. This dataset includes over 9,000 images from the SONIC dataset [38]. The study that recorded this dataset examined nevi in children. The resulting images contain a skin lesion and often large colorful patches. The ISIC 2017 and ISIC 2019 challenge datasets contain such images. Hence, the selected models could associate colorful patches with the classes "benign nevi" or "not melanoma".

Classifiers should not base their predictions on any of the features described in this section. They are indicators of biases introduced in the data acquisition process.

5. Results

Tables 1, 2, and 3 contain the results for the three experiments. The rows of these tables are split into three parts: the Perez *et al*. [29] models for melanoma classification, the Perez *et al*. [29] models for seborrheic keratosis classification, and the Gessert *et al*. [14] ensemble of EfficientNets. Since the tests do not agree, we primarily consider majority votes among them and report whether at least two of three tests indicate the use of a feature.

Table 1 contains the results of the validation experiment. All tests indicate that the classifiers do not use the skin lesion's orientation (0 / 23). The majority vote implies the usage of a feature only twice for all of the features and classifiers in Table 1 (2 / 92). Both times it is the Gessert *et al*. [14] model using the image ID feature. The results in Table 1 match the expectations formulated in Section 4.2 that the features are not or very rarely used by classifiers.

Table 2 contains the results of the experiment concerning the ABCD rule features. Dermatologists designed the ABCD rule to distinguish between melanoma and non-melanoma. The first two columns in Table 2 resemble this fact. All Perez *et al*. [29] models trained to identify melanoma use both the asymmetry and the border features (both 9 / 9). In contrast, no model classifying seborrheic keratosis uses these features (both 0 / 9) according to the majority vote. The ensemble of Gessert *et al*. [14] uses only the border feature (6 / 8) but does not rely on the asymmetry (0 / 8). Only three of the melanoma classifier by Perez *et al*. [29] and three of the classifiers of Gessert *et al*. [14] use the color feature (6 / 23). Most of the seborrheic keratosis models of Perez *et al*. [29] use the dermoscopic structures (4 / 6). In contrast, the other two groups of classifiers rarely use this feature (1 / 9 and 1 / 8). Additionally, we observe a stark difference between the PC-test and the HSIC-test regarding the dermoscopic structures feature. While the former test is positive only four times, the latter is positive 20 times.

Table 3 contains the results for known biases. The majority of models use the patient's age (20 / 23). Only four of the models incorporate the sex (4 / 23). For the skin color feature, the results differ among the three groups of classifiers. Only two out of the nine melanoma classifiers of Perez *et al*. [29] rely on this feature (2 / 9). All the seborrheic keratosis classifiers use the skin color (6 / 6). The majority of the classifiers of Gessert *et al*. [14] also incorporate this feature (6 / 8). Finally, we investigate the colorful patches and find that all but one classifier use this feature (22 / 23).

6. Conclusions

In this work, we use the conditional dependence method described in [35] to analyze state-of-the-art skin lesion classifiers by Perez *et al*. [29] and Gessert *et al*. [14].

We validated that the method [35] produces meaningful results. It does not or very rarely indicate the use of features that contain little to no information about the skin lesion. The features we used for validation have a similar complexity as the relevant features. We say that a classifier uses a feature if the majority of tests indicate usage.

Regarding the features named in the ABCD rule, we found that the melanoma classifiers of Perez *et al*. [29] use the asymmetry and border features. The corresponding seborrheic keratosis classifiers use the dermoscopic structure but do not rely on the other features. The classifiers of Gessert *et al*. [14] use the border feature. Note that the color feature was not used extensively by any group of classifiers. This fact might express an inductive bias in deep neural networks of shape over color information. Furthermore, regarding the dermoscopic structures, we found that the nonlinear HSIC detects an influence. However, the linear PC test indicates that the feature does not speak for any class. Hence, the existence of many different dermoscopic structures is likely to make a sample difficult to classify but not more likely to be classified as any specific class. In summary, we found that the state-of-the-art classifiers use some of the ABCD rule features. While this should inspire some trust, more work is needed to ensure that the classifiers make medically sound decisions. We encourage authors to conduct similar experiments on their classifiers.

We further found that the classifiers use bias variables, namely the age and skin color of a patient and the existence of colorful patches in the images. These observations are worrying, and more work is needed before we can employ automatic classifiers for diagnoses. A solution to this challenge could be the use of adversarial debiasing strategies during training, such as [32]. Nevertheless, we have to be aware that datasets might contain many unknown biases.

References

[1] International skin imaging collaboration, ISIC Archive. https://www.isic-archive.com/.

[2] *Coefficient of Determination*, pages 88–91. Springer New York, New York, NY, 2008.

[3] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307*, 2018.

[4] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (De) Constructing Bias on Skin Lesion Datasets. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2766–2774, Long Beach, CA, USA, June 2019. IEEE.

[5] Titus J. Brinker, Achim Hekler, Alexander H. Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, Jochen S. Utikal, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54, May 2019.

[6] 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.

[7] M Emre Celebi, Hassan A Kingravi, Bakhtiyar Uddin, Hitoshi Iyatomi, Y Alp Aslandogan, William V Stoecker, and Randy H Moss. A methodological approach to the classification of dermoscopy images. *Computerized Medical imaging and graphics*, 31(6):362–373, 2007.

[8] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

[9] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv:1902.03368 [cs]*, Mar. 2019. arXiv: 1902.03368.

[10] 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.

[11] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy. BCN20000: Dermoscopic Lesions in the Wild. *arXiv:1908.02288 [cs, eess]*, Aug. 2019. arXiv: 1908.02288.

[12] 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, Feb. 2017. Number: 7639 Publisher: Nature Publishing Group.

[13] Alan C. Geller, Susan M. Swetter, Katie Brooks, Marie-France Demierre, and Amy L. Yaroch. Screening, early detection, and trends for melanoma: current status (2000-2006) and future directions. *Journal of the American Academy of Dermatology*, 57(4):555–572; quiz 573–576, Oct. 2007.

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

[15] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[16] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer, 2007.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, Dec. 2015. arXiv: 1512.03385.

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]*, Jan. 2018. arXiv: 1608.06993.

[19] Reda Kasmi and Karim Mokrani. Classification of malignant melanoma and benign skin lesions: implementation of automatic abcd rule. *IET Image Processing*, 10(6):448–455, 2016.

[20] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[21] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[24] Nabin K. Mishra and M. Emre Celebi. An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning. *arXiv:1601.07843 [cs, stat]*, Jan. 2016. arXiv: 1601.07843.

[25] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Cnn fixations: an unraveling approach to visualize the discriminative image regions. *IEEE Transactions on Image Processing*, 28(5):2116–2125, 2018.

[26] Sherin Muckatira. Properties Of Winning Tickets On Skin Lesion Classification. *arXiv:2008.12141 [cs, eess]*, Aug. 2020. arXiv: 2008.12141.

[27] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. 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, Apr. 1994.

[28] Judea Pearl. *Causality*. Cambridge university press, 2009.

[29] Fábio Perez, Cristina Vasconcelos, Sandra Avila, and Eduardo Valle. Data augmentation for skin lesion analysis. In *Proceedings of the Third ISIC Workshop on Skin Image Analysis*, pages 303–311. Springer, 2018.

[30] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[31] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

[32] Christian Reimers, Paul Bodesheim, Jakob Runge, and Joachim Denzler. Towards Learning an Unbiased Classifier from Biased Data via Conditional Adversarial Debiasing. *arXiv:2103.06179 [cs]*, Mar. 2021. arXiv: 2103.06179.

[33] Christian Reimers and Christian Requena-Mesa. Deep learning–an opportunity and a challenge for geo-and astrophysics. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, pages 251–265. 2020.

[34] Christian Reimers, Jakob Runge, and Joachim Denzler. Using causal inference to globally understand black box predictors beyond saliency maps. In *International Workshop on Climate Informatics (CI)*, 2019.

[35] Christian Reimers, Jakob Runge, and Joachim Denzler. Determining the relevance of features for deep neural networks. In *European Conference on Computer Vision*, pages 330–346, 2020.

[36] Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *arXiv:1909.13584 [cs, stat]*, Oct. 2020. arXiv: 1909.13584.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[38] Alon Scope, Michael A. Marchetti, Ashfaq A. Marghoob, Stephen W. Dusza, Alan C. Geller, Jaya M. Satagopan, Martin A. Weinstock, Marianne Berwick, and Allan C. Halpern. The study of nevi in children: Principles learned and implications for melanoma diagnosis. *Journal of the American Academy of Dermatology*, 75(4):813–823, Oct. 2016.

[39] 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.

[40] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1143–1151, 2015.

[41] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[42] Wilhelm Stolz and Michael Kunz. Abcd rule — dermoscopedia. https://dermoscopedia.org/w/index.php?title=ABCD_rule&oldid=15572, 2021. [Online; accessed 1-February-2021].

[43] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv:1602.07261 [cs]*, Aug. 2016. arXiv: 1602.07261.

[44] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:1905.11946 [cs, stat]*, Sept. 2020. arXiv: 1905.11946.

[45] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7):938–947, 2019.

[46] 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, Aug. 2018. Number: 1 Publisher: Nature Publishing Group.

[47] Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. Deep neural network or dermatologist? *arXiv:1908.06612 [cs, eess, stat]*, 11797:48–55, 2019. arXiv: 1908.06612.

[48] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks (2013). *arXiv preprint arXiv:1311.2901*, 2013.

[49] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.