Investigating the ABCDE Rule in Convolutional Neural Networks

Federico Bolelli, Luca Lumetti, Kevin Marchesini,
Ettore Candeloro, and Costantino Grana

Università degli Studi di Modena e Reggio Emilia, Modena, Italy
{name.surname}@unimore.it

Abstract. Convolutional Neural Networks (CNNs) have been broadly employed in dermoscopic image analysis, mainly due to the large amount of data gathered by the International Skin Imaging Collaboration (ISIC). But where do neural networks look? Several authors have claimed that the ISIC dataset is affected by *strong biases*, *i.e.,* spurious correlations between samples that machine learning models unfairly exploit while discarding the useful patterns they are expected to learn. These strong claims have been supported by showing that deep learning models maintain excellent performance even when "no information about the lesion remains" in the debased input images. With this paper, we explore the interpretability of CNNs in dermoscopic image analysis by analyzing which characteristics are considered by autonomous classification algorithms. Starting from a standard setting, experiments presented in this paper gradually conceal well-known crucial dermoscopic features and thoroughly investigate how CNNs performance subsequently evolves. Experimental results carried out on two well-known CNNs, EfficientNet-B3, and ResNet-152, demonstrate that neural networks autonomously learn to extract features that are notoriously important for melanoma detection. Even when some of such features are removed, the others are still enough to achieve satisfactory classification performance. Obtained results demonstrate that literature claims on *biases* are not supported by carried-out experiments. Finally, to demonstrate the generalization capabilities of state-of-the-art CNN models for skin lesion classification, a large private dataset has been employed as an additional test set.

Keywords: ABCDE Rule · Convolutional Neural Networks · Skin Lesion Classification · Dataset Bias · Transfer Learning

1 Introduction

Skin cancer is the most common form of human cancer and a major public health issue. Malignant melanoma, although less common, is responsible for most of the deaths [6]. The early detection of skin cancer remains one of the key factors in preventing its progression to advanced stages and lowering mortality rates [40]. To do so, many dermatologists rely on dermoscopy, which is a form of in-vivo skin surface microscopy performed using special equipment to enhance the visibility of the pigmentation of the lesion and perform a faster, more accurate diagnosis over time. Unfortunately, dermoscopy image analysis must be performed by expert clinicians to be effective, and this is why many efforts have been made toward the creation of tools to assist non-specialized physicians in the analysis of dermoscopic images [2]. The outstanding results of deep learning in many different research areas [5,23,26,47], make it one of the most employed and effective options for analyzing medical images. However, the great discriminative power of neural networks comes at the cost of very low explainability. Hence, it is extremely difficult to understand the reasoning behind a model prediction [22,38], and this characteristic can also lead to the possibility of CNNs learning a *bias*. A bias can exist in different shapes and forms and may originate from different sources [35,46], but in the analysis carried out in this paper, we focus on *data-to-algorithm* biases, which, when used by machine-learning training algorithms, might result in biased algorithmic outcomes. In particular, a *dataset bias* can be defined as a collection of features that are semantically irrelevant to the investigated task, but which can be (undesirably) exploited by neural networks to improve the evaluation metrics, hindering their generalization capabilities [32]. This phenomenon has been thoroughly investigated by several authors [4,20] and our goal is to explore it in dermoscopic image analysis [19]. It is desirable for automatic skin lesion classification algorithms to focus on medically relevant features instead of considering irrelevant artifacts (*e.g.,* checkerboard patterns introduced by sharpening filters, black round borders, pen drawings, rulers, and hair) which should be ignored for classification.

The most common dermoscopic relevant features for melanoma detection, explicitly outlined by expert practitioners, are defined in the so-called "ABCDE rule": lesion *Asymmetry*, *Border* irregularity, *Color* variegation, *Diameter* (> 6 mm), and *Evolution* over time [40].

Our study investigates how the performance of CNNs correlates with established dermoscopic criteria by methodically altering images to omit each of the ABCDE melanoma indicators. By selectively "removing" these elements, the research aims to discern the extent to which CNNs rely on authentic clinical features versus incidental image attributes. The contributions of this paper can be summarized as follows:

i) Making use of state-of-the-art interpretability tools, we examine the correlation between deep learning algorithms and well-known dermoscopic features (ABCDE rule) used by expert practitioners to perform diagnoses;
ii) We propose an extensive set of experiments to highlight how the discriminative power of state-of-the-art CNNs is affected by different dermoscopic features and verify the literature claims on dermoscopic datasets biases;
iii) We validate the generalization capabilities of state-of-the-art CNN algorithms for skin lesion classification by carrying out experiments on two totally distinct datasets: the combined ISIC2019 and ISIC2020 and a privately owned one that has no intersection with the former.

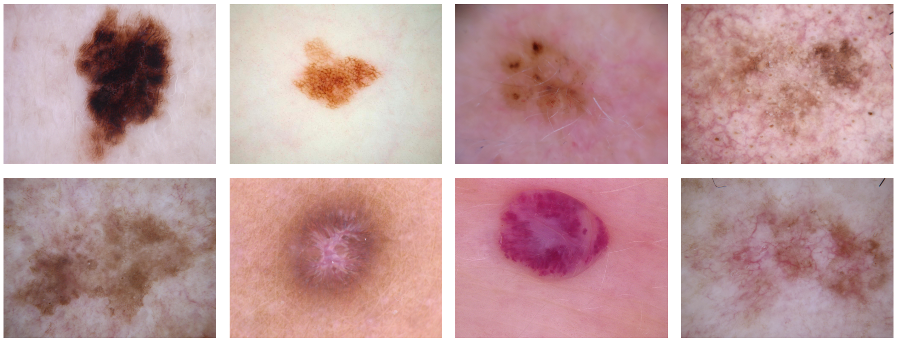


Fig. 1. Samples of the 2019 ISIC dataset. From left to right, top to bottom: Melanoma, Melanocytic Nevus, Basal Cell Carcinoma, Actinic Keratosis, Benign Keratosis, Dermatofibroma, Vascular Lesion, and Squamous Cell Carcinoma.

2 Related Work

CNNs have become the dominant machine learning approach, and the *scaling up* strategy [24] has been widely used to achieve accuracy results similar to those of dermatologists, particularly in skin lesion classification [1,16,17,27,33,34,52,39] and to aid in diagnosis even on low-resolution non-dermoscopic images [15]. However, despite their success, concerns about CNN focusing on irrelevant artifacts were highlighted by [31,41], where studies on multiple COVID-19 datasets, performed hiding sensitive information with large black squares, showed state-of-the-art networks focusing on dataset-specific features rather than clinically relevant ones, highlighting the incompatibility of those models for clinical usage.

In dermatology, Bissoto *et al.* [4] showed the effects of performing skin lesion classification while occluding the actual skin lesion with large black bounding boxes, obtaining a melanoma/non-melanoma classification AUC (Area Under the ROC Curve) score of 77.4%, which is quite inferior compared to state-of-the-art methods, but higher than what expert dermatologists can do [7], highlighting a potential reliance on non-diagnostic features. Additional studies confirmed the CNNs' learned filters focusing on both relevant features (*e.g.,* borders, and colors) and extraneous features (like artifacts surrounding the lesion) [3,54].

Autonomous systems in medical applications aim to act as support tools for clinicians and, therefore, must be trustworthy and highly interpretable. To aid in this task, the outcome explainability of neural networks can be increased thanks to several visualization strategies, like *CAM* (Class Activation Mapping), which have been proposed for the identification of image regions that most contribute to the final prediction [43,49].

In this paper, we make use of state-of-the-art interpretability tools, along with quantitative results, to examine the correlation between deep learning algorithms and well-known dermoscopic features [29] introduced in Section 1.

Table 1. Class distribution of the three employed datasets: 2019 and 2020 ISIC datasets and private dataset.

Class	Label	ISIC2019 %	ISIC2020 %	Private %
Melanoma	MEL	17.8	1.8	16.7
Melanocytic Nevus	NV	50.8	15.7	58.1
Basal Cell Carcinoma	BCC	13.0	–	7.6
Actinic Keratosis	AK	3.0	–	1.6
Benign Keratosis	BKL	10.0	0.7	6.3
Dermatofibroma	DF	0.9	–	1.0
Vascular Lesion	VASC	1.0	–	0.0
Squamous Cell Carcinoma	SCC	2.4	–	1.8
Unknown	unknown	–	81.9	6.9
Total		25 331	33 126	25 849

3 Dermoscopic Images

ISIC. The International Skin Imaging Collaboration (ISIC) began to aggregate a large-scale, publicly available collection of dermoscopic skin lesion images (Fig. 1) starting in 2016, with the aim of supporting research towards enhancing machine learning algorithms for automated skin cancer analysis, showcasing the results of researchers in several challenges and workshops hosted over the years [14]. The 2019 version of the ISIC archive contains a total amount of 25 331 labeled dermoscopy images, belonging to nine different classes [25], which represent eight types of skin lesion plus an additional category, not available in the training partition and containing dermoscopic images of different natures with respect to the other eight classes.

The available data is heavily imbalanced in classes, therefore the 2019 challenge official metric was the balanced accuracy, computed as the average sensitivity among classes regardless of their occurrence in the test set.

The successive 2020 SIIM-ISIC challenge dataset [42] gained patient-level contextual information, providing for each image an identifier that allows lesions from the same patient to be mapped to one another. This additional knowledge is frequently used by clinicians to diagnose melanoma and is especially useful in ruling out false positives in patients with many atypical nevi, leveraging the "ugly duckling sign" rule [18]. The challenge edition, hosted on Kaggle,[1] switched to a binary classification problem: benign or malignant, employing the AUC evaluation metric. In the subsequent sections of the paper, the name ISIC19-20 will be used to refer to the combination of ISIC2019 and ISIC2020 datasets. More details about such a combination are provided in Section 5. Table 1 summarizes ISIC dataset features.

Private Dataset. In order to evaluate the generalization capabilities of state-of-the-art CNNs models, we extend the experiments by means of a private dermoscopic dataset (Fig. 2) consisting of 25 849 images, collected between 2003 and 2019 in the University Hospital of Modena using several distinct acquisi-

[1] kaggle.com/c/siim-isic-melanoma-classification

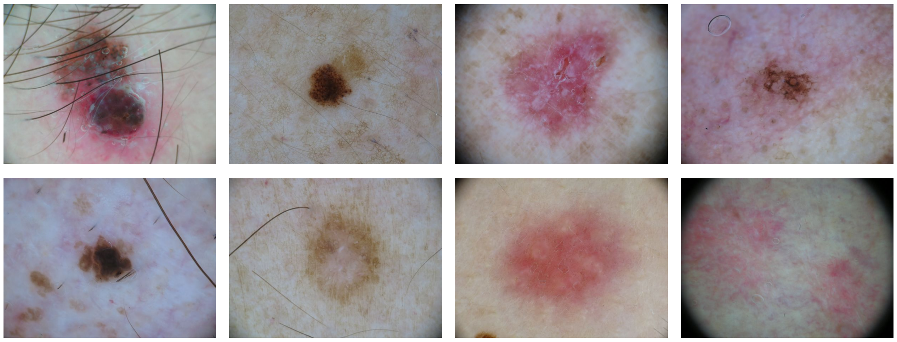


Fig. 2. Samples of the Private dataset. From left to right, top to bottom: Melanoma, Melanocytic Nevus, Basal Cell Carcinoma, Actinic Keratosis, Benign Keratosis, Dermatofibroma, Vascular Lesion, and Squamous Cell Carcinoma.

tion tools, and employing the same classes mapped into the ISIC2019 dataset.[2] This dataset presents a different category distribution compared to the ISIC2020 dataset, with a higher percentage of melanoma cases (Table 1). Similar to both ISIC datasets, the private collection of data contains several clinical information such as sex, age, and site of the lesion. Contrary to the public ISIC dataset, visual artifacts that could be considered a source of biases, such as rulers, ink markings/staining, and colored patches, are almost completely absent in our private dataset (7% ruler, 1.9% ink, and no images with patches). The whole set of dermoscopic images is used as an additional test set for the experiments and analyses carried out in this paper, and thus yields important information about the generalizability of state-of-the-art CNNs models and their possible application in real-world scenarios.

4 Investigating ABCDE Features

Neural networks for skin lesion classification have been shown to focus on relevant features for dermoscopic image analysis, aligned with the ABCDE rule [29,40], but they might also focus on irrelevant visual aspects that are common in malignant skin lesion images, such as artifacts related to pen drawings, markers, colored patches, or rulers. Moreover, additional research showed that CNNs are able to recognize acquisition device models and calibration settings, thus identifying the provenience of an image that might be highly related to the final diagnosis [31]. Hence, it is extremely important to be able to interpret which image characteristics neural networks take into account when making a class prediction to highlight potential *data-to-algorithm* biases. This can be achieved

[2] The dataset is currently under review by the ethical committee to be publicly released. After approval, it will be accessible at https://ditto.ing.unimore.it/.

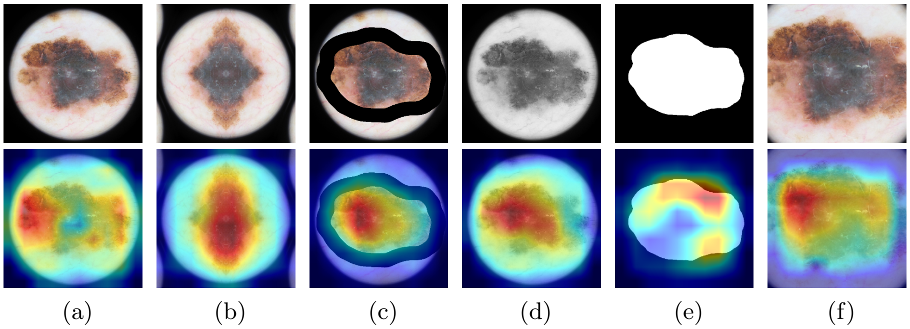


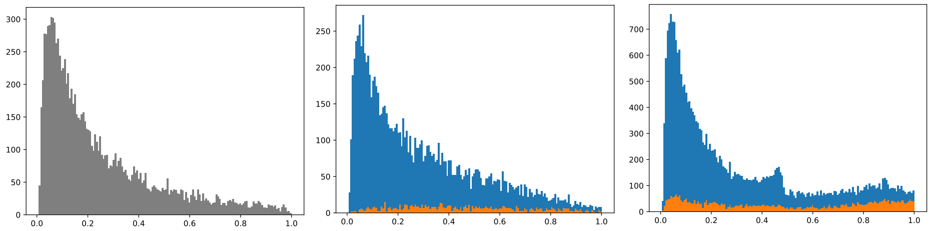
Fig. 3. Grad-CAM visualization when debasing different ABCD(E) properties. (a) Original, (b) Asymmetry, (c) Borders, (d) Grayscale, (e) Mask, (f) Diameter

by means of Class Activation Mapping (CAM) strategies, employed in this paper. In particular, Grad-CAM [43] was exploited to locate the regions of an image that most contribute to the final prediction. We run an extensive set of experiments to study how introducing noise in the ABCDE properties affects neural network performance and analyze which sections of an image CNNs focus on when crucial features are debased or removed. Some of the experiments described in the following exploit segmentation masks obtained by means of DeepLabv3+ [12], trained using the 2017 ISIC segmentation task dataset [9]. Sample images obtained through feature debasing are reported in Fig. 3 and generated by means of the European Computer Vision Library (ECVL) [10]. Additional examples of ABCDE features debasing images can be found in Fig. 7 at the end of the paper.

By applying the feature debasing process described in the following of this section, we obtain five additional variations of each considered dataset (*i.e.,* five variations of the ISIC datasets and five variations of the Private dataset), each of them is employed for both training (ISIC19-20) and testing (ISIC19-20 and private dataset) selected models.

Tampering with Asymmetry. Asymmetry is one of the most important visual features for melanoma detection [36], it can be described as the difference in volume and shape of two parts of a skin lesion, obtained by *cutting* it with a straight line passing through its center. In order to train a symmetry-agnostic neural network, dermoscopic images can be split by a random straight line and by its perpendicular, both passing through the center of the lesion. Subsequently, a quarter of the image can be flipped over both axes to obtain a version of the original lesion with increased symmetry. Practically, the center of the image is aligned with the centroid of the lesion obtained from the segmentation mask. The image is then randomly rotated, and the top-right quarter is flipped with respect to the horizontal and vertical axes (Fig. 3b).

Concealing Borders. With the aim of removing valuable information about the shape of a skin lesion edge, which is a crucial aspect when assessing its malignancy, we cover borders with a thick black line obtained from the contour of the segmentation map. Firstly, a morphological dilation operation is applied



(a) (b) (c)

Fig. 4. Histograms of foreground density distribution within different test sets. (a) ISIC2020 official test set, (b) ISIC19-20 "Internal" test set, (c) Private Dataset. Benign and malignant skin lesions are depicted in blue and orange, respectively. Best viewed in color.

to the contour, with a kernel size proportional to both the image size and the foreground-background ratio. Then, to smooth out irregular segments, a Gaussian filter with a large kernel is applied. Finally, the black border image is superimposed on the original one, thus removing any information about the actual transition from the human skin (background) to the actual lesion (foreground). Fig. 3c showcases an example of the end result image.

Removing Color. The presence of multiple colors within a single mole (blue, black, white, red, and brown) or the uneven distribution of color can sometimes be a warning sign of melanoma, since most benign lesions are usually a single shade of brown or tan [30]. Two different sets of experiments are conducted in order to assess the effects of discarding information about color from dermoscopic images. The first one is run by simply converting the image from RGB to grayscale, thus removing any knowledge about the different colors within a skin lesion (Fig. 3d). However, while this processing step erases any data about hue and saturation, it does not affect the luminance, thus leaving the CNN the chance to learn valuable features from the color distribution within moles. In the second experiment, color features are completely removed as we train a neural network to classify skin lesions using only their segmentation masks (Fig. 3e). In this extreme setup, the neural network is fed with minimal knowledge about skin lesions, and is forced to make a prediction based uniquely on noisy, automatically obtained, binary mole shapes.

Altering Diameter. Because skin cancer cells grow abnormally fast, diameter is one of the most important parameters in skin lesion classification. Unfortunately, dermoscopic images are acquired at several scales, which are not always included as metadata and can not be deducted from the image, as only a limited amount of samples contain a ruler. As a matter of fact, a mole that exceeds the borders of the image is not necessarily larger than one that does not; information about diameter is extremely noisy and very hard to investigate, yet potentially extremely important for melanoma detection algorithms. To remove any information about mole dimensions, the foreground-background ratio of images is set to a fixed percentage. We choose to train a CNN uniquely with samples where the skin lesion represents the 80% of the image, as it yields good qualitative results. In order to achieve this, samples where the skin lesion is contained in just a small portion of the image must be cropped, using the mole centroid as the center, whereas images with moles covering more than 80% of the original sample are padded by reflecting the sections of the image closest to the borders. An example of the result of this process is illustrated in Fig. 3f. Additionally, the foreground density histograms in Fig. 4 show that this method mostly results in crops, whereas only a very small portion of the dataset (foreground percentage $> 80\%$) needs padding. By following the aforementioned *Crop&Pad* technique, in the rare cases of very elongated lesions, a small part of the mole is left out by the crop. However, each image will present the same number of foreground pixels, thus eliminating the image scale differences and bringing all the lesions to the same size.

About Evolving. Dermoscopic images seldom contain information about evolution, as only in a few cases follow-up data is provided in the existing datasets. Introducing such additional data in dermoscopic datasets would certainly have significant implications in the research field, but it cannot be considered and analyzed nowadays. For this reason, we were unable to experiment on lesion evolution.

5 Experiments

Datasets Preprocessing. To harmonize the two ISIC datasets, the following pre-processing steps have been employed: first, the 2020 classes are mapped into 2019 ones; then, in order to compensate for dissimilarities between image sizes, a squared center crop is performed to produce images of $min(h,w) \times min(h,w)$ pixels, later resized to 768×768. The combined ISIC2019 and ISIC2020 datasets, purged of duplicates [51], provide a total of 57 964 images. As previously mentioned, we refer to this combined dataset as ISIC19-20. As introduced in Section 3, we also employed a private dataset with the same class mapping as the ISIC19-20.

Networks and Training Details. Our study utilizes two of the networks constituting the ensemble strategy adopted by the ISIC2020 Kaggle challenge winner [21], *i.e.,* EfficientNet-B3 and ResNet-152. While achieving performances that are comparable with the state-of-the-art, they have a limited computational load in terms of time and memory and allow us to perform the extensive set of experiments described in this section. Input image sizes are 300×300 and 256×256 for the two models, respectively. Both networks are trained with the Cross-Entropy loss and Adam optimizer [28], with a learning rate of 3×10^{-5}. Networks are trained for 20 epochs and produce 9 class probabilities as output, among which only the melanoma class is considered.

Given the unavailability of ISIC test set labels, we expanded our evaluation metrics by partitioning the validation set of ISIC19-20 to create an "internal" test set: the resulting dataset counts 46 379 training images, 1 159 for validation,

Table 2. Experimental results obtained by training (and testing) the models on the input configurations described in Section 4. Each of the *Experiment* correspond to a training performed on the corresponding debased ISIC19-20 dataset and tested on the debased ISIC19-20 "internal" test set. Threshold is set to 0.5.

Model	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy
EfficientNet-B3	Original	0.9671	0.7821	0.7180	0.9808	0.7487	0.9577
	Asymmetry	0.9448	0.7755	0.5399	0.9850	0.6366	0.9459
	Borders	0.9605	0.7326	0.6678	0.9766	0.6987	0.9495
	Color (Grayscale)	0.9559	0.7420	0.7071	0.9763	0.7241	0.9527
	Color (Mask)	0.8017	0.6897	0.0656	0.9972	0.1198	0.9154
	Diameter	0.9724	0.8216	0.7399	0.9845	0.7786	0.9631
ResNet-152	Original	0.9572	0.7548	0.6934	0.9782	0.7228	0.9531
	Asymmetry	0.9188	0.6539	0.4848	0.9837	0.5568	0.9320
	Borders	0.9456	0.7548	0.6043	0.9706	0.6699	0.9475
	Color (Grayscale)	0.9424	0.7216	0.5788	0.9784	0.6424	0.9432
	Color (Mask)	0.8502	0.6073	0.1136	0.9206	0.1914	0.9154
	Diameter	0.9553	0.7688	0.6513	0.9811	0.7052	0.9520

Table 3. Experimental results obtained by training (and testing) the models on the input configurations described in Section 4. Each of the *Experiment* correspond to a training performed on the corresponding debased ISIC19-20 dataset and tested on the debased private dataset. Threshold is set to 0.5.

Model	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy
EfficientNet-B3	Original	0.7983	0.5299	0.5038	0.9104	0.5165	0.8425
	Asymmetry	0.7693	0.5553	0.4025	0.9354	0.4667	0.8465
	Borders	0.7896	0.5261	0.4992	0.9099	0.5123	0.8413
	Color (Grayscale)	0.7673	0.4607	0.4540	0.8935	0.4573	0.8201
	Color (Mask)	0.7032	0.6017	0.0322	0.9957	0.0612	0.8349
	Diameter	0.8099	0.5597	0.5168	0.9185	0.5374	0.8515
ResNet-152	Original	0.7872	0.4774	0.5542	0.8772	0.5129	0.8229
	Asymmetry	0.7340	0.5279	0.3176	0.9416	0.3966	0.8351
	Borders	0.7559	0.4498	0.4921	0.8762	0.4700	0.8107
	Color (Grayscale)	0.6860	0.3565	0.4411	0.8389	0.3943	0.7719
	Color (Mask)	0.6881	0.5243	0.1187	0.8436	0.1936	0.8313
	Diameter	0.7660	0.4121	0.5424	0.8409	0.4684	0.7899

and 10 426 images for testing. The *private dataset* is employed for testing classification performance as well. These datasets, modified as outlined in Section 4 facilitated a broader analysis across five variant datasets, against which designated architectures are trained and tested. Table 2 and Table 3 report results obtained by training the model on the debased ISIC19-20 datasets and testing on the ISIC19-20 "internal" test set and on the private dataset, respectively.

6 Discussion

The efficacy of our ABCDE feature-concealment methods, despite occasional inaccuracies in segmentation mask generation (Fig. 5), underscores their ability to divert neural networks' focus from compromised features towards other ones, as demonstrated in most test scenarios (Fig. 3).

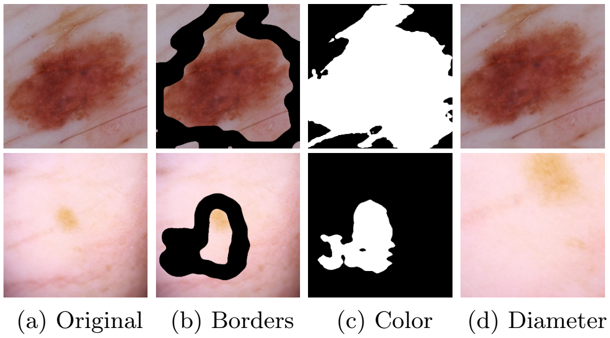


Fig. 5. Example of failure cases due to wrongly generated masks.

In particular, in Fig. 3c the neural network trained to classify images with hidden borders makes a prediction focusing on the section of the lesion with the most variance of color intensity. The same patch is equally important for the network fed with grayscale images (Fig. 3d), whereas Fig. 3a shows that mole borders are of great interest for the “standard” model.

On the other hand, CNN asked to make a prediction based solely on the segmentation mask strictly focuses on the sections of the foreground with higher concavity, which is roughly the only *valuable* piece of information about the lesion to be found in the extremely degraded input.

As suggested by the high classification performance, neural networks autonomously learn to extract features for melanoma detection. The accuracy obtained when single important features are missing is very close to the “reference” values, meaning that the other image features are enough to produce a satisfying classification prediction. Experimental results alone are clearly not enough to distinguish whether such features are biases or notoriously important elements for melanoma detection. For this reason, in our work, we also rely on the clinically validated Grad-CAM analysis (Fig. 3).

The discriminative power of CNNs is also confirmed on the private dataset. CNN performance tends to drop when the source domain (training data) and the target domain (test data) come from distinct origins, even for extremely simple tasks such as handwritten digit recognition, where classification accuracy across separate datasets can be decreased by up to 40% [50]. A performance drop can be due to a large number of reasons (biases), such as different lighting settings, resolution, image quality, human-introduced artifacts, subject centering, and image acquisition devices [53,37].

Notably, this is also confirmed by our experiments, identifying that model generalization abilities are satisfactory (AUC performance is higher than those obtained by expert dermatologists [7]), but certainly require fine-tuning the models on the real case scenario they have to be employed, thus ensuring an adequate level of Precision and Recall.

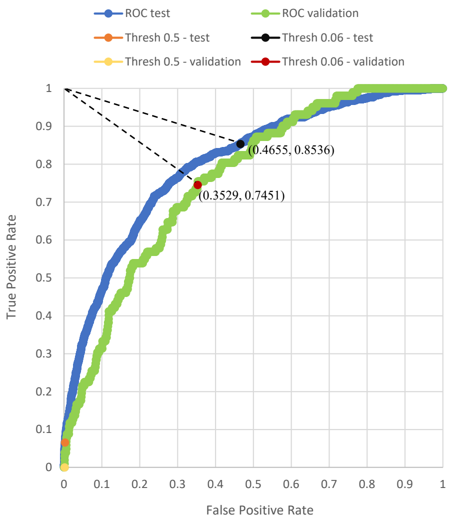


Fig. 6. ROC curves for the *Mask* experiment on the ISIC19-20 "internal" validation and test sets (Table 2 and Table 5 of the paper). The threshold value that minimizes the distance from the (0;1) are highlighted in both ROC curves. Moreover, the points corresponding to the 0.5 threshold are highlighted in the two curves.

About AUC. The Area Under the Receiver Operating Characteristic curve (AUC) is a well-known metric designed to evaluate the diagnostic capabilities of binary classifiers. It is the official metric of the ISIC2020 challenge, and it offers the advantage of not needing a fixed threshold, thus supplying one less parameter to "overfit" proposed algorithms on the official test set. However, real clinical applications require a threshold to be set and a class prediction to be given; evaluating experimental results uniquely using the AUC metric can be misleading. To put results into context, we further discuss the performance of the CNN trained to classify skin lesion binary masks (*i.e.,* debased dataset obtained removing colors). Focusing on the EfficientNet-B3 results in Table 2, the fifth line shows that the investigated network yields an AUC of 0.8017 when tested on the subset of public images (ISIC19-20) used as an "internal" test set. This is the area under the blue curve in Fig. 6. When following this strategy, we obtain a tool with a sensitivity of 0.0656, and a specificity of 0.9972, which means that the model can correctly recognize only 6.5% of the melanoma cases, but successfully identifies 99.7% of the not-melanoma cases. Clearly, this is not the positive result that an AUC of 0.8017 might suggest.

As a matter of fact, the assumption that 0.5 is an appropriate threshold when dealing with neural networks is not correct [41], as shown in Fig. 6. Alternatively, the threshold can be set by studying the ROC curve obtained on the validation

Table 4. Experimental results obtained by training (and testing) the models on the input configurations described in Section 4. Each of the *Experiment* correspond to a training performed on the corresponding debased ISIC19-20 dataset and tested on the debased ISIC19-20 "internal" test set using a specific threshold calculated as the value of the ROC curve which minimizes the distance from (0;1) on the validation set.

Model	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy
EfficientNet-B3	Original	0.9671	0.3163	0.9519	0.8020	0.4748	0.8152
	Asymmetry	0.9448	0.2527	0.9628	0.7260	0.4003	0.7468
	Borders	0.9605	0.2218	0.9858	0.6672	0.3621	0.6952
	Color (Grayscale)	0.9559	0.2590	0.9628	0.7349	0.4082	0.7549
	Color (Mask)	0.8017	0.1500	0.8536	0.5345	0.2551	0.5625
	Diameter	0.9724	0.2419	0.9803	0.7044	0.3881	0.7287

Table 5. Experimental results obtained by training (and testing) the models on the input configurations described in Section 4. Each of the *Experiment* correspond to a training performed on the corresponding debased ISIC19-20 dataset and tested on the private test set using a specific threshold calculated as the value of the ROC curve which minimizes the distance from (0;1) on the validation set.

Model	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy
EfficientNet-B3	Original	0.7983	0.2578	0.8570	0.5055	0.3963	0.5642
	Asymmetry	0.7693	0.2140	0.9017	0.3365	0.3460	0.4308
	Borders	0.7896	0.2012	0.9300	0.2600	0.3308	0.3718
	Color (Grayscale)	0.7673	0.2291	0.8844	0.4038	0.3639	0.4840
	Color (Mask)	0.7032	0.2421	0.7604	0.5229	0.3672	0.5626
	Diameter	0.8099	0.2150	0.9263	0.3221	0.3489	0.4230

Table 6. Experimental results using foreground densities obtained from segmentation masks, bounding boxes, and bounding boxes that cover at least 70% of the image as melanoma probability on the ISIC19-20 "internal" test set and on the private dataset.

Dataset	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Acc.
ISIC19-20 "Internal" test set	Segm. Mask	0.7215	0.1483	0.7388	0.5917	0.2470	0.6046
	B. Box	0.7154	0.1483	0.7202	0.6019	0.2459	0.6123
	B. Box 70%	0.6220	0.1830	0.3989	0.8286	0.2509	0.7909
Private dataset	Segm. Mask	0.6980	0.2856	0.5898	0.7043	0.3848	0.6852
	B. Box	0.6919	0.2573	0.6589	0.6190	0.3701	0.6256
	B. Box 70%	0.6517	0.3328	0.4735	0.8098	0.3909	0.7536

set (green curve in Fig. 6), and choosing the value in the graph closer to point $(0;1)$, *i.e.,* the value that maximizes *True Positive Rate* while minimizing *False Positive Rate*. In this particular case, the desired rate is ≈0.06, and by employing this same threshold on the EfficientNet-B3 CNN outputs over the test set, we obtain a binary classifier with a sensitivity of 0.8536 and a specificity of 0.5345. Table 4, Table 5, and Table 6 present the results obtained by setting the prediction threshold following the described steps, always making use of the validation set. Regardless of how thresholds are set, it is clear that high AUC values do not always correspond to satisfying discriminative capabilities.

Finally Identifying the "Bias" in Dermoscopic Datasets. Contrasting with Bissoto et al.'s findings [4] where CNNs performed well despite significant lesion occlusion, our analysis suggests that lesion size can be inferred from the foreground-background ratio and significantly influences predictions. While Bissoto et al. observed high AUCs (0.712) with major lesion coverage ($\geq$ 70%), our evaluations posit that networks might rely on lesion dimensions rather than intricate pixel patterns unrelated to the mole. This is substantiated by our *Segmentation Mask* and *Bounding Box* experiments (Table 6), where AUCs correlate strongly with lesion area metrics, even without deep learning models. This experiment has been pushed further by making predictions based only on lesion bounding box (and not segmentation mask) dimensions and, finally, by setting the foreground-background ratios as $\geq 70\%$. Results obtained are reported in the aforementioned table with the name of *Bounding Box* and *Bounding Box 70%*. Finally, histograms in Fig. 4 show that the probability of a lesion being malignant grows with its size within a dermoscopic image. Intuitively, a mole that exceeds the borders or gets very close to them is not necessarily larger than others, but it is more likely to be malignant. This characteristic might be more related to the complexity of including a whole malignant lesion when acquiring dermoscopic images [13,44], than to the diameter itself. Nevertheless, this feature is strongly related to the nature of dermoscopic images, and experiments provided in [4] are insufficient to prove the presence of biases in the ISIC dataset.

7 Conclusion

In this work, we explored the correlation between automatic skin lesion classification and the ABCDE rule. This was done by gradually removing important visual information from CNN inputs and analyzing performance changes. Experimental results show that neural networks autonomously learn to extract features that are notoriously important for melanoma detection, but also prove that their performance is still satisfying when some of these features are removed. Our experiments provide *no proof* that this is related to dataset biases: instead, the remaining information can be enough to achieve satisfying or even good classification accuracy. As pointed out by different authors [45,48], the interpretation of GradCAM's saliency maps may be subjective to reader biases and cannot be used to draw general conclusions about network behavior. However, combined with the quantitative evaluation discussed and showcased in this paper, they contribute to our final conclusion.

In particular, the proposed paper experimentally proved that the foreground-background ratio is strongly related to the malignancy probability of a skin lesion. The reasoning behind this might be related to the well-known *diameter* characteristic from the ABCDE rule, but also to the fact that capturing the entire malignant mole in a dermoscopic image is usually not trivial given its dimensions, the non-clearly defined borders, and the irregular shapes that characterize cancerous skin lesions [8,11,13,44]. Nevertheless, foreground-background ratio is a valuable dermoscopic property. We cannot conclude that "there are no biases in the ISIC dataset", but we can certainly state that literature claims of strong biases affecting the ISIC dataset are supported by an inconsistent experimental analysis.

Finally, testing model performance on a totally distinct private dataset, with no possible intersection with samples employed during the training phase, demonstrated that, despite intra-datasets biases (if any), state-of-the-art algorithms preserve satisfactory performance: still higher than those obtained by expert dermatologists [7], but with lower Precision and Recall.

Acknowledgements. This work was supported by the University of Modena and Reggio Emilia and Fondazione di Modena, through the FAR 2023 and FARD-2023 funds (Fondo di Ateneo per la Ricerca).

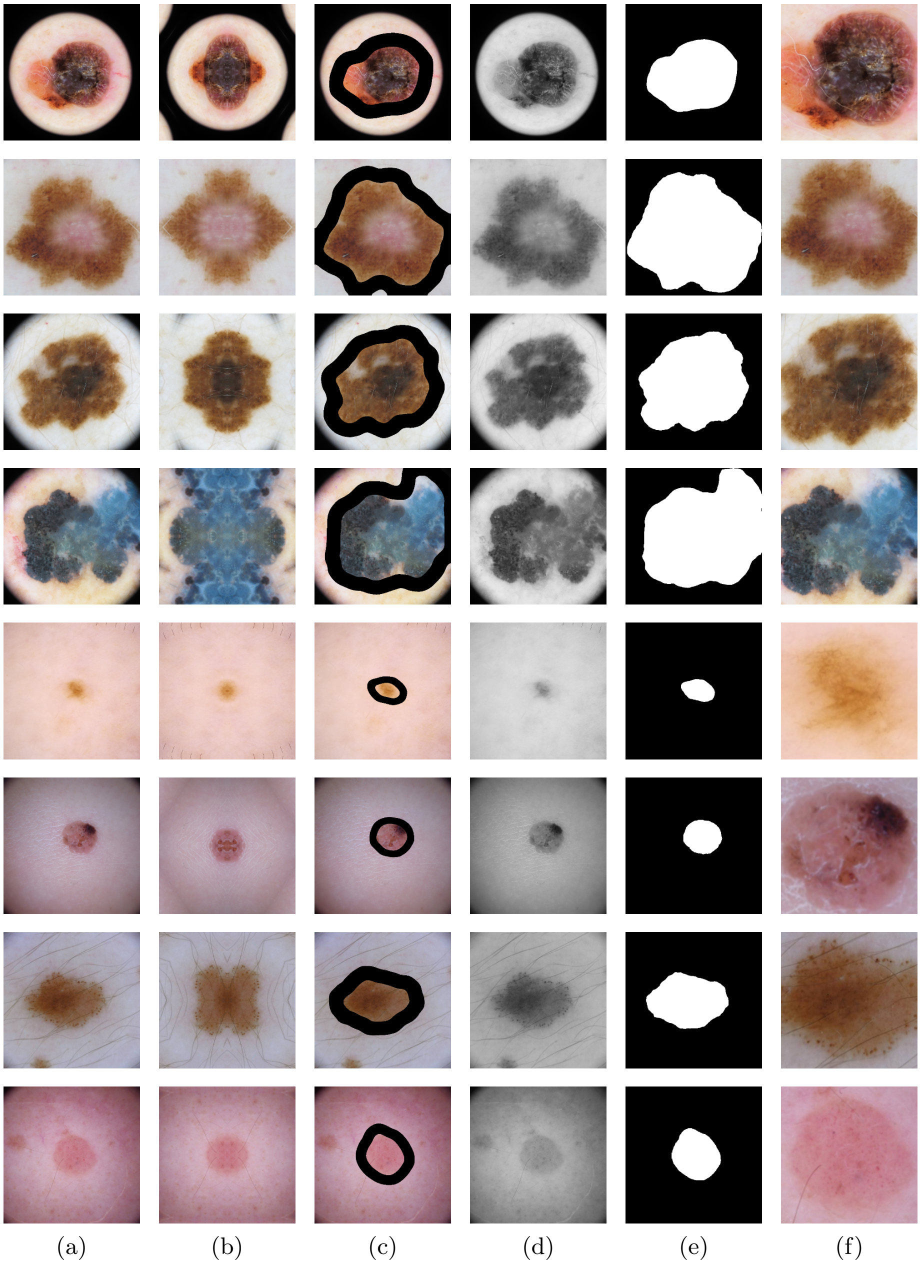


Fig. 7. Skin lesion image samples obtained from ISIC dataset after debasing different ABCDE properties. Columns from left to right: (a) Original, (b) Asymmetry, (c) Borders, (d) Color - Grayscale, (e) Color - Mask, (f) Diameter. The first half rows depict melanomas, while the others are generated from benign lesions.

References

1. Abayomi-Alli, O.O., Damasevicius, R., Misra, S., Maskeliunas, R., Abayomi-Alli, A.: Malignant skin melanoma detection using image augmentation by oversamplingin nonlinear lower-dimensional embedding manifold. Turkish Journal of Electrical Engineering and Computer Sciences **29**(8) (2021)
2. Allegretti, S., Bolelli, F., Pollastri, F., Longhitano, S., Pellacani, G., Grana, C.: Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021)
3. Barata, C., Celebi, M.E., Marques, J.S.: Explainable Skin Lesion Diagnosis Using Taxonomies. Pattern Recognition **110** (2020)
4. Bissoto, A., Fornaciali, M., Valle, E., Avila, S.: (De)Constructing Bias on Skin Lesion Datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
5. Bolelli, F., Baraldi, L., Grana, C.: A Hierarchical Quasi-Recurrent approach to Video Captioning. In: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS) (2018)
6. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **68**(6) (2018)
7. Brinker, T.J., Hekler, A., Hauschild, A., Berking, C., Schilling, B., Enk, A.H., Haferkamp, S., Karoglan, A., von Kalle, C., Weichenthal, M., Sattler, E., Schadendorf, D., Gaiser, M.R., Klode, J., Utikal, J.S.: Comparing artificial intelligence algorithms to 157 german dermatologists: the melanoma classification benchmark. European Journal of Cancer **111** (2019)
8. Brú, A., Albertos, S., Subiza, J.L., García-Asenjo, J.L., Brú, I.: The Universal Dynamics of Tumor Growth. Biophysical journal **85**(5) (2003)
9. Canalini, L., Pollastri, F., Bolelli, F., Cancilla, M., Allegretti, S., Grana, C.: Skin Lesion Segmentation Ensemble with Diverse Training Strategies. In: International Conference on Computer Analysis of Images and Patterns. Springer (2019)
10. Cancilla, M., Canalini, L., Bolelli, F., Allegretti, S., Carrión, S., Paredes, R., Gómez, J.A., Leo, S., Piras, M.E., Pireddu, L., Badouh, A., Marco-Sola, S., Alvarez, L., Moreto, M., Grana, C.: The DeepHealth Toolkit: A Unified Framework to Boost Biomedical Applications. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021)
11. Capdehourat, G., Corez, A., Bazzano, A., Alonso, R., Musé, P.: Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions. Pattern Recognition Letters **32**(16) (2011)
12. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Proceedings of the European conference on computer vision (ECCV) (2018)
13. Claridge, E., Hall, P., Keefe, M., Allen, J.: Shape analysis for classification of malignant melanoma. Journal of Biomedical Engineering **14**(3) (1992)
14. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., 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)

15. Di Biasi, L., De Marco, F., Auriemma Citarella, A., Castrillón-Santana, M., Barra, P., Tortora, G.: Refactoring and performance analysis of the main cnn architectures: using false negative rate minimization to solve the clinical images melanoma detection problem. BMC Bioinformatics **24** (2023)
16. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639) (2017)
17. Fujisawa, Y., Otomo, Y., Ogata, Y., Nakamura, Y., Fujita, R., Ishitsuka, Y., Watanabe, R., Okiyama, N., Ohara, K., Fujimoto, M.: Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. British Journal of Dermatology **180**(2) (2019)
18. Gaudy-Marqueste, C., Wazaefi, Y., Bruneu, Y., Triller, R., Thomas, L., Pellacani, G., Malvehy, J., Avril, M.F., Monestier, S., Richard, M.A., et al.: Ugly Duckling Sign as a Major Factor of Efficiency in Melanoma Detection. JAMA Dermatology **153**(4) (2017)
19. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut Learning in Deep Neural Networks . Nature Machine Intelligence **2**(11) (2020)
20. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (2019)
21. Ha, Q., Liu, B., Liu, F.: Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge. arXiv preprint 2010.05351 (2020)
22. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A.: Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation **16**(1), 45–74 (2024)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
25. Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N.C.F., Rotemberg, V., Halpern, A.C., Reiter, O., Carrera, C., Barreiro, A., Helba, B., Puig, S., Vilaplana, V., Malvehy, J.: BCN20000: Dermoscopic Lesions in the Wild. Sci. Data (2024)
26. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine **29**(6) (2012)
27. Kadry, S., Taniar, D., Damaševičius, R., Rajinikanth, V., Lawal, I.A.: Extraction of Abnormal Skin Lesion from Dermoscopy Image using VGG-SegNet. In: 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII) (2021)
28. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR). San Diega, CA, USA (2015)

29. Lattoofi, N.F., Al-Sharuee, I.F., Kamil, M.Y., Obaid, A.H., Mahidi, A.A., Omar, A.A., et al.: Melanoma Skin Cancer Detection Based on ABCD Rule. In: 2019 First International Conference of Computer and Applied Sciences (CAS). IEEE (2019)
30. Lynn, N.C., Kyu, Z.M.: Segmentation and Classification of Skin Cancer Melanoma from Skin Lesion Images. In: 2017 18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT). IEEE (2017)
31. Maguolo, G., Nanni, L.: A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images. arXiv preprint 2004.12823 (2020)
32. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys (CSUR) **54**(6) (2021)
33. Mondal, B., Das, N., Santosh, K., Nasipuri, M.: Improved Skin Disease Classification Using Generative Adversarial Network. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE (2020)
34. Nawaz, M., Nazir, T., Masood, M., Ali, F., Khan, M.A., Tariq, U., Sahar, N., Damaševičius, R.: Melanoma segmentation: A framework of improved DenseNet77 and UNET convolutional neural network. International Journal of Imaging Systems and Technology (2022)
35. Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data **2** (2019)
36. Pellacani, G., Grana, C., Seidenari, S.: Algorithmic reproduction of asymmetry and border cut-off parameters according to the ABCD rule for dermoscopy. Journal of the European Academy of Dermatology and Venereology **20**(10) (2006)
37. Perone, C.S., Ballester, P., Barros, R.C., Cohen-Adad, J.: Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. NeuroImage **194** (2019)
38. Pollastri, F., Maroñas, J., Bolelli, F., Ligabue, G., Paredes, R., Magistroni, R., Grana, C.: Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021)
39. Pollastri, F., Parreño, M., Maroñas, J., Bolelli, F., Paredes, R., Ramos, D., Grana, C.: A Deep Analysis on High Resolution Dermoscopic Image Classification. IET Computer Vision (2021)
40. Rigel, D.S., Russak, J., Friedman, R.: The Evolution of Melanoma Diagnosis: 25 Years Beyond the ABCDs. CA: A Cancer Journal for Clinicians **60**(5) (2010)
41. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L., et al.: Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence **3**(3) (2021)
42. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data **8**(1) (2021)
43. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
44. Senan, E.M., Jadhav, M.E.: Analysis of dermoscopy images by using ABCD rule for early detection of skin cancer. Global Transitions Proceedings **2**(1) (2021)
45. Srinivas, S., Fleuret, F.: Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability. In: International Conference on Learning Representations (2021)

46. Suresh, H., Guttag, J.: A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In: Equity and access in algorithms, mechanisms, and optimization. Association for Computing Machinery (2021)
47. Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: International Conference on Machine Learning (2019)
48. Viviano, J.D., Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Saliency is a Possible Red Herring When Diagnosing Poor Generalization. In: International Conference on Learning Representations (2021)
49. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2020)
50. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312** (2018)
51. Weber, J.: True duplicates in ISIC 2020 dataset (2020), https://www.kaggle.com/c/siim-isic-melanoma-classification/discussion/161943
52. Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., Zhao, S.: Skin Cancer Classification With Deep Learning: A Systematic Review. Frontiers in Oncology (2022)
53. You, K., Long, M., Cao, Z., Wang, J., Jordan, M.I.: Universal Domain Adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
54. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting Deep Visual Representations via Network Dissection. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(9) (2019)