Investigating the ABCDE Rule: Are Skin Lesion Datasets Really Biased?

Federico Bolelli, Luca Lumetti, Kevin Marchesini, Ettore Candeloro, and Costantino Grana
Università degli Studi di Modena e Reggio Emilia, Italy
{name.surname}@unimore.it

1. Skin Cancer & CAD Systems

Skin cancer is a major public health issue, with **melanoma** causing most of the deaths.

Early detection with dermoscopy (a form of surface microscopy) is fundamental to lower mortality rates but **requires expertise** [1].

Many efforts have been made to create computer-aided diagnosis (**CAD**) **systems** to assist non-specialized clinicians in early **detection of skin cancer** [2].

2. CNNs and Biases

Convolutional Neural Networks (CNNs) employed in modern CAD systems achieve performance comparable to those of dermatologists [3] but present some **challenges**:

- **CNNs lack explainability**, hiding possible *biases*.
- CNNs may rely on irrelevant dataset features, hindering their generalization abilities [4].
- Existing studies indicate CNNs can maintain high performance even when lesions are occluded, suggesting **potential data-to-algorithm biases** [5].

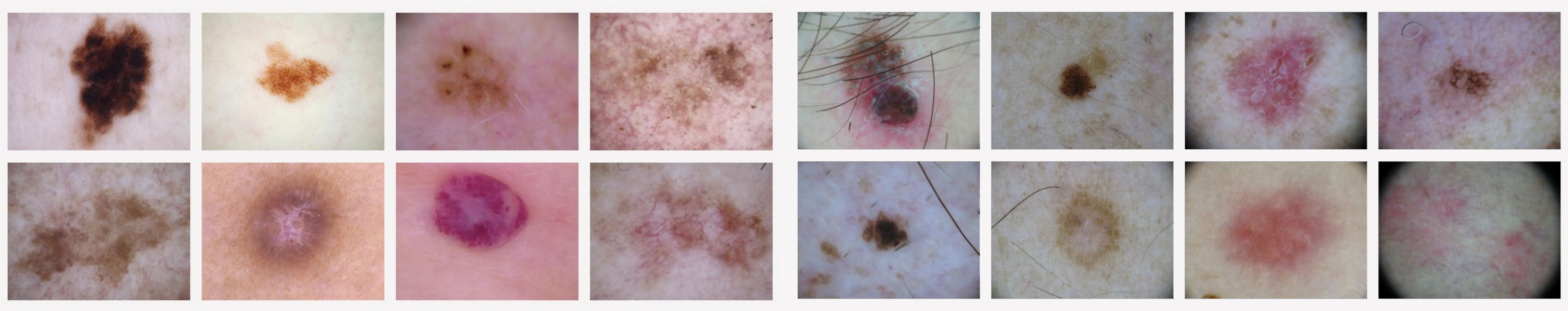
Clinical Focus: CNNs should prioritize clinically relevant features, such as those defined in the **ABCDE rule.**

Objective: to study how CNN performance correlates with dermoscopic criteria by **removing ABCDE** skin cancer **features** from the data and **evaluating potential biases**.

3. Dataset

Two skin lesion datasets with no overlapping were employed for our experiments. Dataset classes were transformed into **tumor vs benign lesion labels** for binary classification:

- **SIIM-ISIC 2019-2020**: composed of 57.964 dermoscopic images of nine different types of skin lesions, collected from 2016 in various hospitals [6].
- **PRIVATE**: composed of 25.849 dermoscopic images of nine different types of skin lesions, collected between 2003 and 2019 in the University Hospital of Modena.

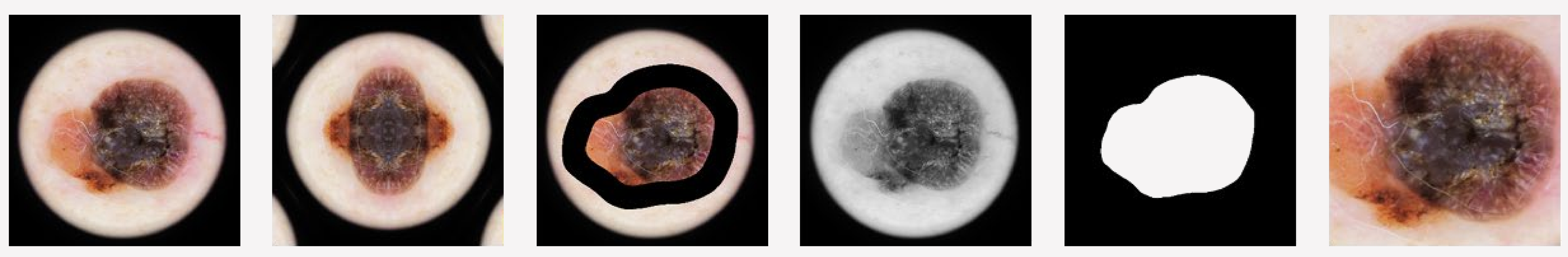


SIIM-ISIC 2019 + 2020 PRIVATE

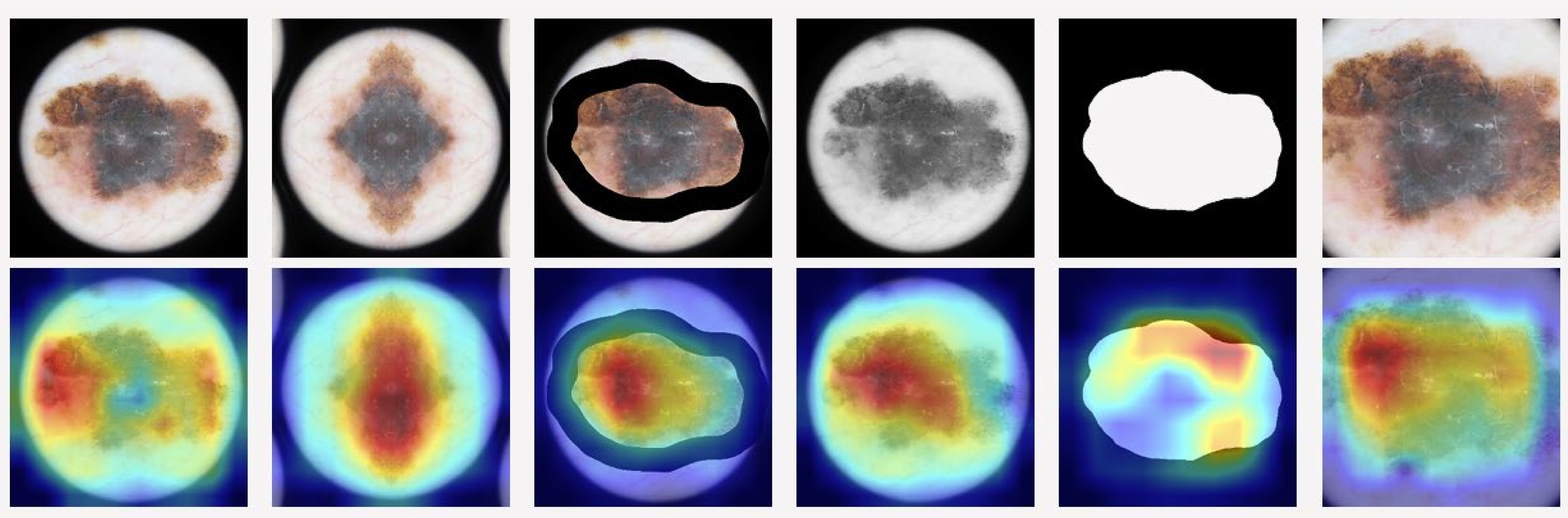
4. ABCDE Feature Debasing

Dermoscopic features (ABCDE rule) were systematically **altered on both datasets:**

- **Asymmetry**: lesions modified to be symmetrical;
- **Borders**: edges were concealed with black masks;
- **Color**: converted to grayscale or replaced with mask;
- **Diameter**: lesion relative size normalized within images;
- **Evolution**: not considered due to limited temporal data.



Grad-CAM activations were studied to analyze the visual features that debased trained CNNs relied upon when classifying skin lesions.



5. Experiments on Covering Lesions

Contrary to previous work [5], our results show that **CNN performance is linked to lesion dimensions**, with malignancy predictions increasing as lesion size grows. This indicates **CNNs rely on lesion size over unrelated patterns**, debunking earlier claims of dataset bias.

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

6. Experiments on ABCD Features & Results

Training on a subset of the SIIM-ISIC 19-20, **testing** on the SIIM-ISIC test set and the private dataset, to **evaluate generalization abilities**.

Performances remain satisfactory even when using **debased ABCD features** both for training and testing.

Grad-CAM visualizations confirm **CNNs adapt to alternative relevant features** when other ABCD visual aspects are debased.

Foreground-background ratio correlates with malignancy predictions, suggesting **CNN exploit the lesion size** (**D**iameter in the ABCDE rule) instead of other uncorrelated features.

Conclusion. There is **no proof that CNNs rely on dataset-to-algorithm biases**. Instead, they uses clinically relevant features available in the image to achieve robust classification, even when certain features are debased.

SIIM-ISIC 2019-2020

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

PRIVATE

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

References

[1] 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)
[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] 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)
[4] 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)
[5] 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)
[6] Codella, N.C., Gutman, et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI) (2018)