Universität Augsburg
Fakultät für Angewandte Informatik

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

Sixth ISIC Skin Image Analysis Workshop @CVPR 2021 Virtual

Fabian Stieler, Fabian Rabe, Bernhard Bauer

19.06.2021

SPONSORED BY THE
Federal Ministry of Education and Research

Agenda

Introduction

Skin cancer detection

- Popular application for clinical decision support [4]
- Deep Neural Networks (DNNs) as viable method
  to develop a model for classifying skin images [1, 5, 7, 20]

Model Interpretation

- Increasing attention in AI research
- Recent work has recognized the need for skin image classification tasks [2, 6, 20]
  - DNNs are often considered as black box models
  - Critical for use in safety-relevant environments (medical field)

Need of Explainability

Not only the model itself, but also the explanations have to be adapted
to the problem in order to be useful for the particular use case [11]

Domain Specific Explainable AI

Explainable Artificial Intelligence (XAI)

- Field of research focuses on making a model's predictions understandable
- Many innovative techniques recently emerged [3,8,9,12,19,21]
- Useful in many aspects:
  - Model Debugging,
  - Model Knowledge Extraction,
  - Bias detection, …
- Helping with the interpretation of model (or "AI-system") behavior

Need of <u>customized</u> explanations for specific domains

AI systems are more likely to be accepted by users,
if the results can be explained in a human way. [11]

Domain Specific Explainable AI

Model Interpretation Methods

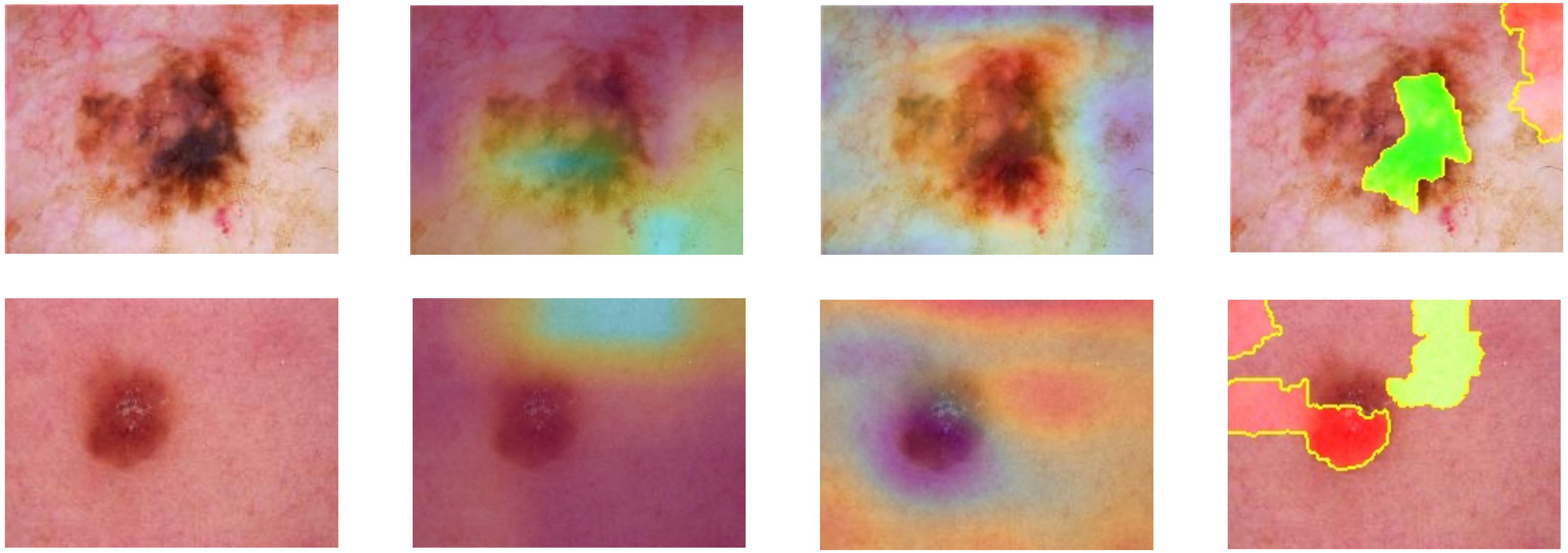
- Post-hoc explanations: Interpretation of predictions from a previously trained black box model
- Local explanations: Individual predictions

ISIC_0025732

True Class: Melanoma
Prediction: Melanoma (0.999)

ISIC_0024362

True Class: Nevus
Prediction: Nevus (0.999)



Original | Grad-CAM [15] | RISE [13] | LIME [14]

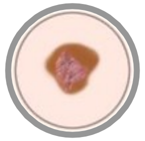
Domain Specific Explainable AI

ABCD rule of dermatoscopy [16]

- Dermatologist's Human Approach for melanoma detection
- A score is calculated using the following properties

Asymmetry

"Irregularly shaped"



Border irregularity

abrupt truncation of the pigment pattern at the border

Color variegation

single vs. multiple colors

Differential Structure

different structural components

- Lesion is examined for all four criteria separately → Score finally leads to a diagnosis

Explainer for Skin Lesion Classifier

Idea: Domain-Specific Explanations by linking LIME with the ABCD-Rule

Local Interpretable Model-agnostic Explanations (LIME) [14]	ABCD rule of dermatoscopy [16]

Local Interpretable Model-agnostic Explanations (LIME) [14]

- Generic ML-Model interpretation method
- Perturbation based
- Suitable for image data

ABCD rule of dermatoscopy [16]

- Easy to understand
- Leads to accurate classifications
- Characteristics used to classify the lesion can be scored independently

Customize LIME's perturbation logic with the criteria of the ABCD rule

Instead of selecting image areas with super-pixels and occluding them, the skin image is modified along diagnostic characteristics.

Explainer for Skin Lesion Classifier

Perturbation Dimensions

Boundary

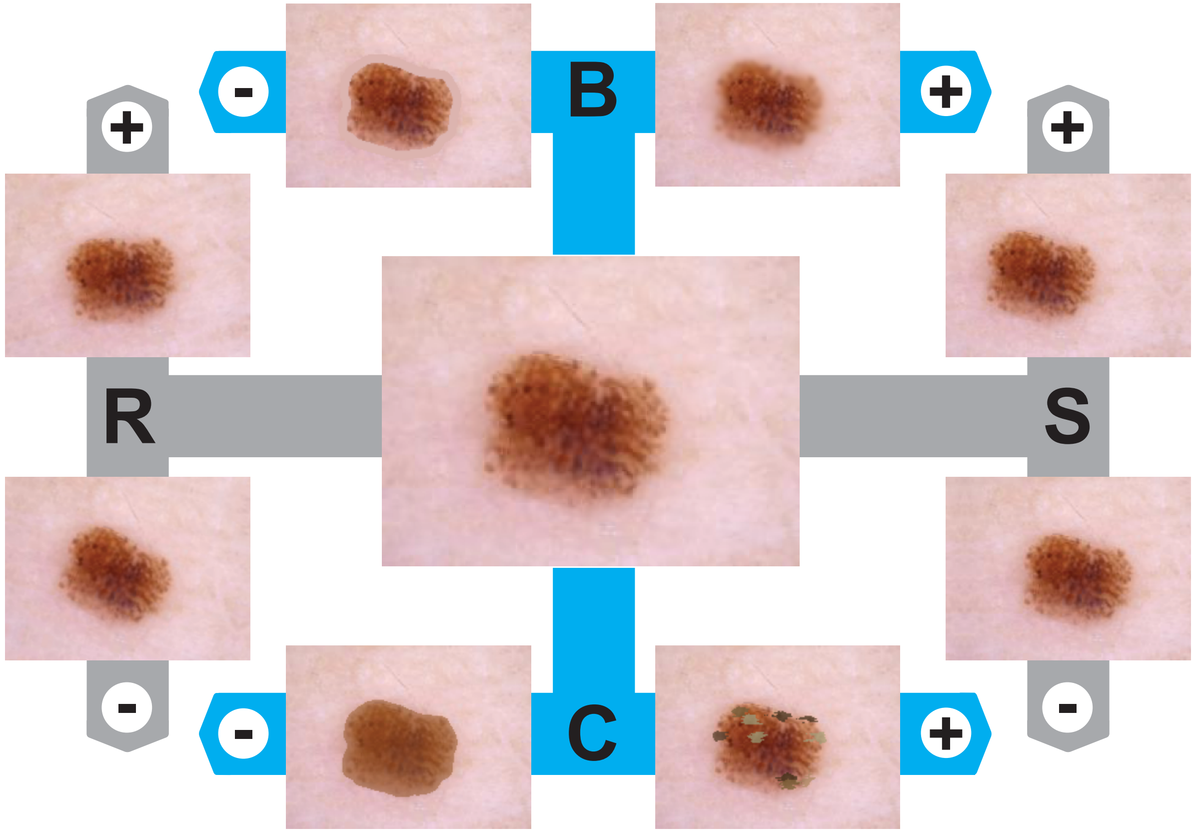
- (-) Sharply delineated line around the lesion
- (+) Gaussian blur is added in edge region

Color

- (-) Lesion segmentation turned into a uniform color
- (+) Add random color patches

Rotate / Shift

- (-/+) Medically irrelevant perturbations



Explainer for Skin Lesion Classifier

Hypothesis

Investigation of a two-class problem allows the derivation of the following hypotheses:

Boundary

- \- Sharply delineated line around the lesion → Prediction for nevus will **decrease**
- \+ Gaussian blur is added in edge region → Prediction for nevus will **increase**

Color

- \- Lesion segmentation turned into a uniform color → Prediction for nevus will **decrease**
- \+ Add random color patches → Prediction for nevus will **increase**

Rotate / Shift

- -/+ Medically irrelevant perturbations → Prediction will **not change**

Experiment and Empirical Results

DNN-based skin image classifier

Model

- Pre-trained MobileNet [10]
- Transfer learning - Approach
- No Data Augmentation,
  no feature engineering
- F_1-Score Nevus: 0.91
- F_1 -Score Melanoma: 0.57
- F_1 -Score Macro average: 0.74

Data

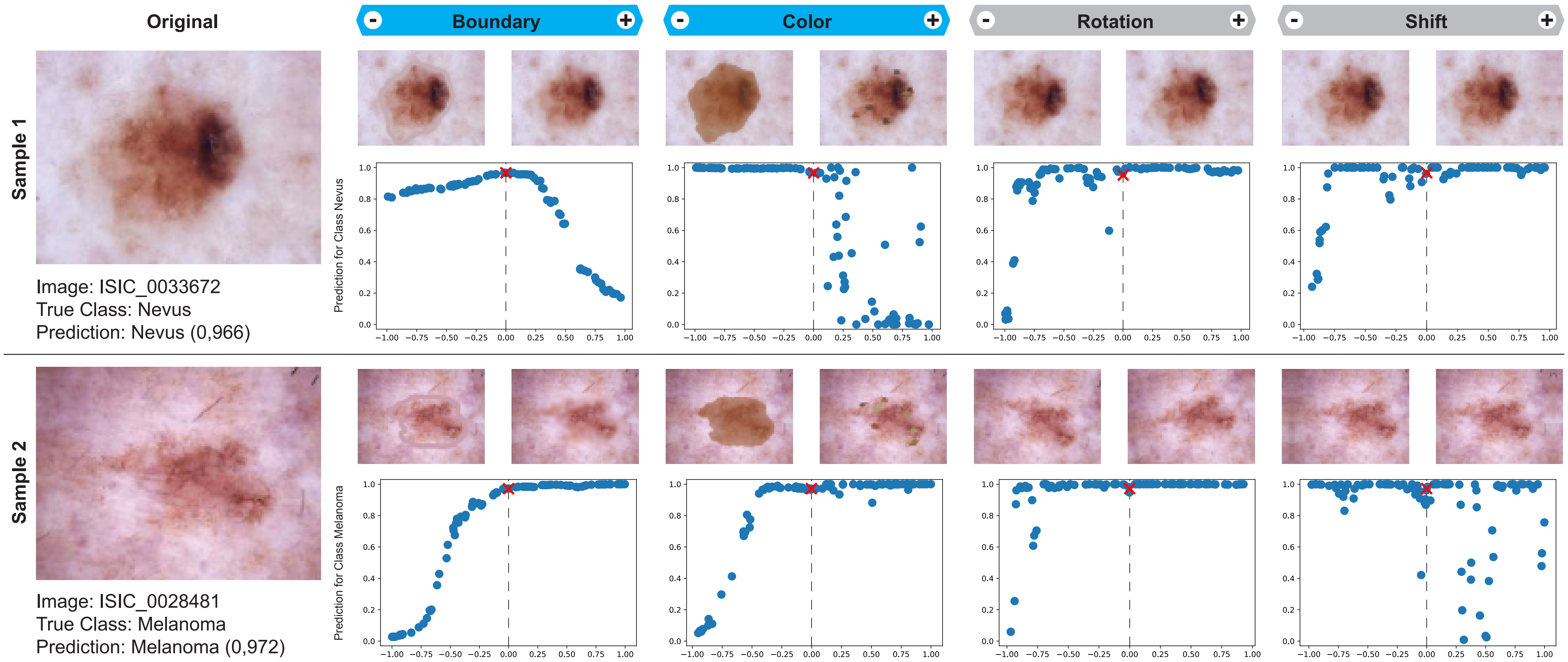
- HAM10000 data set [18]:
  collection of multi-source dermatoscopic images, annotated by dermatosogists
- Segmentation data [17] to select the relevant perturbation-area around the lesion
- 6,705 images of nevi and
  1,113 images of melanoma
- Train/Test Split ratio: 80/20

Explainer Experiment

- Explanations generated on selected input-samples from the test dataset
- For each input-sample, 50 perturbed samples along all dimensions were generated

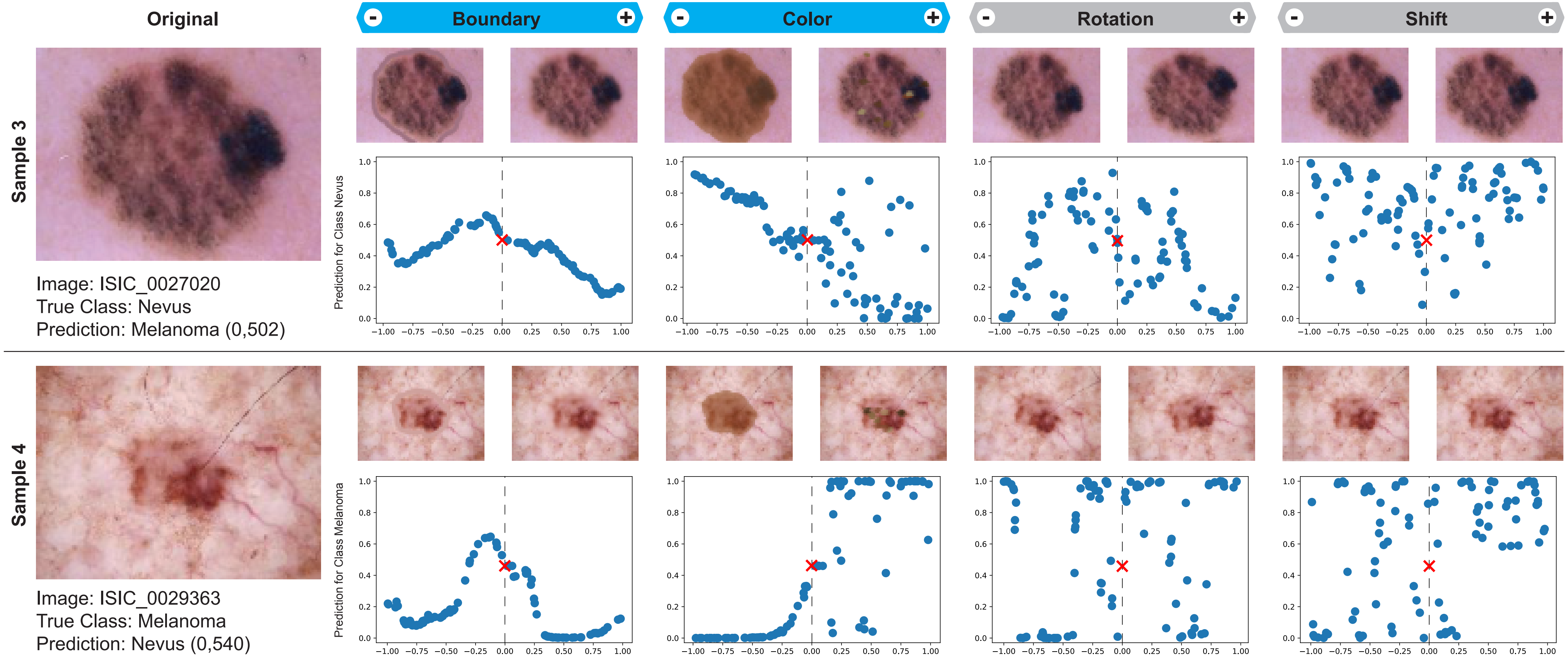
Experiment and Empirical Results

True Positive: “In which dimensions does the model remain accurate?”



Experiment and Empirical Results

False Positive: “Why did the model fail?”



Summary and Outlook

Summary

- Models/Classifier in an AI-based system only provide predictions
  - Physicians and patients can not ask “Why?” the model came to its decision
  - XAI methods are intended to meet this need
  - There is a need of domain-specific explanations
- This work showed the idea of the combination of LIME with the ABCD rule
- Explainer was demonstrated on selected lesion samples

Outlook

- Experiments can be performed with different models and data sets
- Missing evidence between observed importance of feature-dimension and true ABCD score
- Implementation of remaining perturbation dimensions (asymmetry + differential structure)
- The approach of linking a human medical algorithm with an ML-Model interpretation method may be applicable in other medical specialties

References

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

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

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

[4] 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.

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

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

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

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

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

References

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

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

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

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

[14] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should I trust you?”: Explaining the predictions of any classifier. CoRR, abs/1602.04938, 2016.

[15] 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.

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

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

[18] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5(1):180161, 2018.

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

References

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

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