European Journal of Cancer 155 (2021) 191–199

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

A benchmark for neural network robustness in skin cancer classification

Roman C. Maron [a], Justin G. Schlager [b], Sarah Haggenmüller [a], Christof von Kalle [c], Jochen S. Utikal [d,e], Friedegund Meier [f], Frank F. Gellrich [f], Sarah Hobelsberger [f], Axel Hauschild [g], Lars French [b], Lucie Heinzerling [b], Max Schlaak [h], Kamran Ghoreschi [h], Franz J. Hilke [h], Gabriela Poch [h], Markus V. Heppt [i], Carola Berking [i], Sebastian Haferkamp [j], Wiebke Sondermann [k], Dirk Schadendorf [k], Bastian Schilling [l], Matthias Goebeler [l], Eva Krieghoff-Henning [a], Achim Hekler [a], Stefan Fröhling [m], Daniel B. Lipka [n], Jakob N. Kather [o], Titus J. Brinker [a,*]

[a] *Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[b] *Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany*
[c] *Department of Clinical-Translational Sciences, Charité University Medicine and Berlin Institute of Health (BIH), Berlin, Germany*
[d] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[e] *Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[f] *Skin Cancer Center at the University Cancer Centre and National Center for Tumor Diseases Dresden, Department of Dermatology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany*
[g] *Department of Dermatology, University Hospital of Kiel, Kiel, Germany*
[h] *Charité – Universitätsmedizin Berlin, Department of Dermatology, Venereology and Allergology, Berlin, Germany*
[i] *Department of Dermatology, University Hospital Erlangen, Erlangen, Germany*
[j] *Department of Dermatology, University Hospital Regensburg, Regensburg, Germany*
[k] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[l] *Department of Dermatology, Venereology and Allergology, University Hospital Würzburg, Würzburg, Germany*
[m] *Division of Translational Medical Oncology, German Cancer Research Center (DKFZ) & National Center for Tumor Diseases (NCT), Heidelberg, Germany*
[n] *Section of Translational Cancer Epigenomics, Division of Translational Medical Oncology, German Cancer Research Center (DKFZ) & National Center for Tumor Diseases (NCT), Heidelberg, Germany*
[o] *Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany*

Received 26 May 2021; received in revised form 18 June 2021; accepted 29 June 2021
Available online 11 August 2021

* *Corresponding author*: Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany.

E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

https://doi.org/10.1016/j.ejca.2021.06.047

KEYWORDS
Benchmarking;
Dermatology;
Skin neoplasms;
Melanoma;
Nevus;
Artificial intelligence;
Deep learning

Abstract ***Background:*** One prominent application for deep learning–based classifiers is skin cancer classification on dermoscopic images. However, classifier evaluation is often limited to holdout data which can mask common shortcomings such as susceptibility to confounding factors. To increase clinical applicability, it is necessary to thoroughly evaluate such classifiers on out-of-distribution (OOD) data.

Objective: The objective of the study was to establish a dermoscopic skin cancer benchmark in which classifier robustness to OOD data can be measured.

Methods: Using a proprietary dermoscopic image database and a set of image transformations, we create an OOD robustness benchmark and evaluate the robustness of four different convolutional neural network (CNN) architectures on it.

Results: The benchmark contains three data sets—Skin Archive Munich (SAM), SAM-corrupted (SAM-C) and SAM-perturbed (SAM-P)—and is publicly available for download. To maintain the benchmark's OOD status, ground truth labels are not provided and test results should be sent to us for assessment. The SAM data set contains 319 unmodified and biopsy-verified dermoscopic melanoma (n = 194) and nevus (n = 125) images. SAM-C and SAM-P contain images from SAM which were artificially modified to test a classifier against low-quality inputs and to measure its prediction stability over small image changes, respectively. All four CNNs showed susceptibility to corruptions and perturbations.

Conclusions: This benchmark provides three data sets which allow for OOD testing of binary skin cancer classifiers. Our classifier performance confirms the shortcomings of CNNs and provides a frame of reference. Altogether, this benchmark should facilitate a more thorough evaluation process and thereby enable the development of more robust skin cancer classifiers.

1. Introduction

Deep learning (DL)–based computer vision has demonstrated its potential numerous times in the medical field [1]. Because of its amenability to visual pattern recognition, skin cancer classification has especially benefited from increased research interest. Studies have shown that DL-based image classifiers are on par or superior to human experts [2–7,41–43,45,46] or can be used to improve skin cancer diagnostics [8–11]. These results are no longer limited to artificial environments [12,13], and some DL-based skin cancer screening tools have already received market approval [14].

Part of this success can be attributed to the continuous improvement of convolutional neural networks (CNNs), which are at the core of most DL-based image classifiers. However, despite their seemingly superhuman performance, CNN-based image classifiers suffer from a variety of flaws, such as learning of spurious correlations [15], susceptibility to small image changes [16,17] or adversarial vulnerabilities [18,19]. Geirhos *et al.* [20] describe some of these failures as being symptoms of the same underlying problem: shortcut learning. Instead of the classifier learning valid decision rules that generalise to out-of-distribution (OOD) data, such as genuine morphological characteristics of melanomas and nevi, correlating but meaningless features that occur in the data set are frequently learnt (i.e. shortcuts). This often results in non-generalisable or non-robust classifiers. These classifiers usually have good performance on test data that is similar to the training data but perform worse or fail on OOD data [4,21], even if the image changes are just small in nature (e.g. small image rotations or change in brightness) [16,17,22].

Unsurprisingly, skin cancer classifiers also exhibit potential symptoms of shortcut learning such as learning of artifacts [23], adversarial vulnerabilities [24] and general brittleness [22]. OOD testing of skin cancer classifiers should therefore become standard practice. Given the variety of dermoscopic data sets [25–30], OOD testing is certainly possible. However, although these data sets originate from different clinics and presumably present a broad variety of imaging modalities, they might not challenge classifiers sufficiently. In general object recognition for example, the popular ImageNet [31] data set has seen a variety of modifications which are designed to test a classifier's response to various distribution shifts [18,32,33]. However, such challenging test sets are still missing in dermatology.

We therefore introduce the Skin Archive Munich (SAM) data set to further diversify the existing collection of dermoscopic data sets. In addition, we provide two extra data sets derived from SAM—SAM-corrupted (SAM-C) and SAM-perturbed (SAM-P)—which are designed to test skin cancer classifiers against image corruptions and perturbations, respectively. Corruptions test, how well a classifier fares with low-quality images (e.g. blurry, distorted or contaminated), whereas perturbations test classifier stability in response to subtle image changes (e.g. continuous change of image brightness or scale). By doing so, we hope to establish a skin lesion benchmark analogous to the study by Hendrycks and Dietterich [32], which can be used to thoroughly test classifiers in a diverse OOD setting. To provide a frame of reference, we trained and evaluated four common CNN architectures on this benchmark. This study describes the benchmark's structure and the obtained classifier results.

2. Materials and methods

2.1. Study design

We obtained a proprietary database of biopsy-verified dermoscopic melanoma and nevus images which we used to create three data sets. The first data set, referred to as SAM, contains the unmodified images. The second and third data sets, SAM-C and SAM-P, contain artificially corrupted or perturbed images from SAM which were transformed using 22 different types of transformations (see Table 1). Next, we trained four representative CNN architectures and evaluated their performance on this benchmark. For each architecture, training and evaluation were repeated five times to obtain an averaged result less influenced by stochastic training events.

The study design was heavily inspired and adapted from the ImageNet-C/P benchmark study [32]. The Ethics Committee of Heidelberg University waived ethical approval due to the anonymity of the dermoscopic images.

2.2. Unmodified data set—SAM

The SAM data set was compiled at the Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany. It contains dermoscopic images of 85 unique lesions, of which 37 are classified as melanoma and 48 as nevus. The diagnosis for each lesion is histopathologically verified. As each lesion was photographed multiple times at different angles and magnifications, the total number increases to 194 melanoma and 125 nevus images. Lesions were photographed with a NIKON D2X at a resolution of 4288 × 2848 pixels and scaled down to 1000 × 667 pixels to reduce the size of the benchmark without losing too much relevant image information.

2.3. Artificially corrupted data set—SAM-C

The SAM-C data set is designed to test classifier robustness against corruptions. The images are computationally modified to mimic low-quality inputs (e.g.

Table 1
Overview and short description of the applied corruption and perturbation transformations.

Category	Transformation	Description	SAM-C	SAM-P
Noise	Gaussian	Appears in low-lighting conditions.	✓	✓
	Shot	Electronic noise that is caused by the discrete nature of light.	✓	✓
	Impulse	Salt-and-pepper noise caused by bit errors.	✓	✗
	Speckle	Additive noise where noise added to a pixel tends to be larger if the original pixel intensity is larger.	✓✓	✓✓
Blur	Defocus	The image is out of focus.	✓	✗
	Motion	The camera moves quickly in some direction.	✓	✓
	Zoom	The camera moves quickly towards the object.	✓	✓
	Gaussian	A low-pass filter where a blurred pixel is the result of a weighted average of its neighbours. Far away pixels are weighted less.	✓✓	✓✓
Dermoscopy	Black corner	Black image corners caused by the dermatoscope.	✓	✗
	Characters	Letters, numbers and punctuation marks which might be overlaid by the camera.	✓	✓
	Bubbles	Bubbles caused by gel trapped between the skin and the dermatoscope.	✓✓	✓✓
Digital distortions	Brightness up/down	Varies with daylight intensity or the camera.	✓	✓
	Contrast	Depends on lightning conditions and object colour.	✓	✗
	Elastic	Transformations stretching or contracting small image regions.	✓	✗
	Pixelate	Artefacts occurring when upsampling low-resolution images.	✓	✗
	JPEG	Artefacts occurring due to lossy image compression format.	✓	✗
	Saturate	Seen in edited images where images are made more or less colourful.	✓✓	✗
	Translate	Move the image in the horizontal direction.	✗	✓
	Rotate	Rotation in clockwise direction.	✗	✓
	Tilt	Incline at certain angles.	✗	✓
	Scale	Enlargement of the image while maintaining image size.	✗	✓
	Shear	Shift parts of the image to opposite directions.	✗	✓✓

SAM-C, SAM-corrupted; SAM-P, SAM-perturbed.
A total of 22 transformations, drawn from four categories, were applied to SAM-C, SAM-P or both. For SAM-C, the transformation type brightness was applied twice to either increase or decrease image brightness. Double check marks indicate that the transformation is designated as an extra transformation, which is intended to be used for additional experimentation, but which is not included in the final benchmark evaluation.

blurry, obstructed, distorted). The applied corruption types are drawn from four main categories—noise, blur, dermoscopy and digital (see Table 1 and Fig. 1). To mimic varying corruption intensities which could occur in routine clinical care, we apply each corruption type with five different severity levels. We also provide four extra corruption types, one for each category, which are not included in the benchmark evaluation. Instead, these extra corruption types can be used for additional experimentation.

2.4. Artificially perturbed data set—SAM-P

The SAM-P data set is designed to test classifier robustness against perturbations from the same four categories as SAM-C (see Table 1). To create SAM-P, a perturbation sequence containing 31 frames is generated for each image in SAM and for each perturbation type (see Fig. 2). Except for noise sequences, each sequence has temporality, meaning that each frame within a sequence is a perturbation of the previous frame. For noise sequences, the first image is always unmodified and the 30 subsequent images have slight noise perturbations applied to them. Contrary to corruptions, these modifications are not as pronounced and are designed to test a classifier's prediction stability when faced with small image changes. Again, we provide four extra perturbation types to facilitate additional experimentation.

2.5. Models

Pretrained ImageNet models from four different architectures (AlexNet [34], VGG16+BN [35], ResNet50 [36] and DenseNet121 [37]) were trained without data augmentation on images from HAM10000 [25], ISIC [26], BCN20000 [27], PH2 [28], DERM7PT [29] and SKINL2 [30] using fastai [38]. For details, please refer to the Supplementary Methods.

2.6. Evaluation

To obtain a reference of how well a model *f* performs on unmodified images, we first compute its performance on SAM. For this, we calculate the model's clean balanced error rate BE^{f}_{clean} by subtracting its balanced accuracy score obtained on SAM from 1. Next, we assess a model's 'robustness to change' by computing its performance on SAM-C and SAM-P using the following metrics:

SAM-C: We compute each model's balanced error rate $BE^{f}_{s,c}$ on SAM-C for each corruption type *c* and severity level $s (1 \leq s \leq 5)$. We can then get the balanced corruption error BCE^{f}_{c} for a single corruption type *c* by taking the average across all severity levels for that corruption. As presumably not all corruptions are equally difficult, we adjust by a baseline which in our case is AlexNet's balanced error rate $BE^{AlexNet}_{s,c}$. Thus, we get

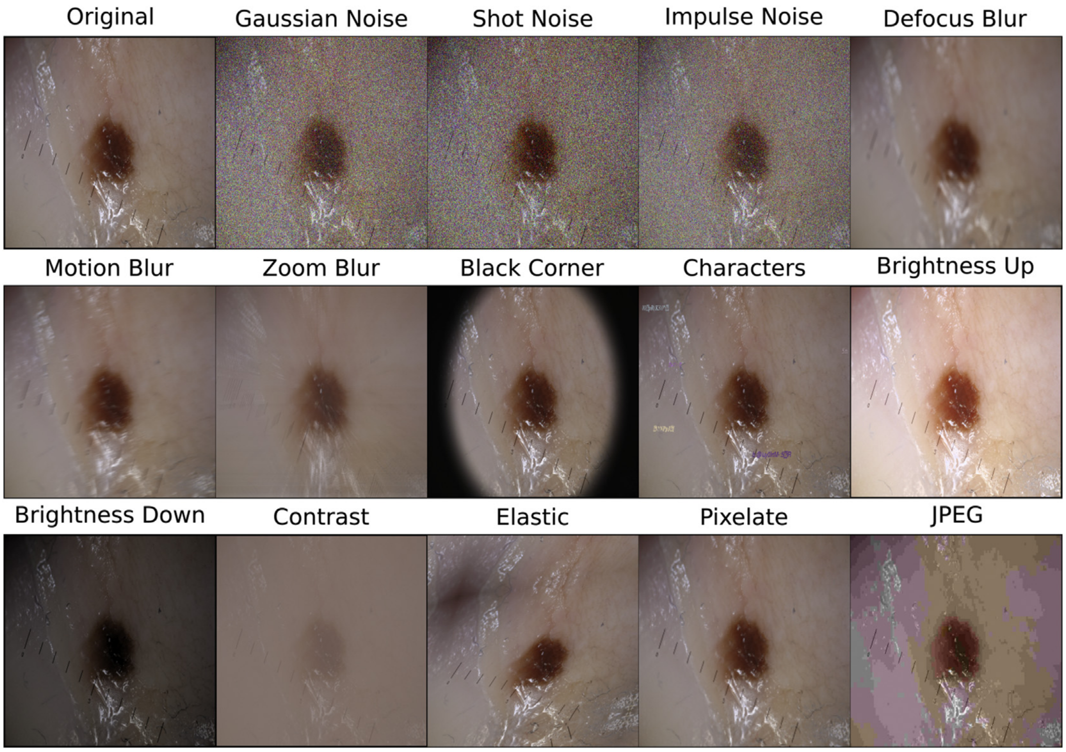

Fig. 1. **Overview of SAM-C corruptions.** Shown here are all corruption types applied to an exemplary dermoscopic image. For reference, we also show the unmodified image in the top left corner. SAM-C, SAM-corrupted.

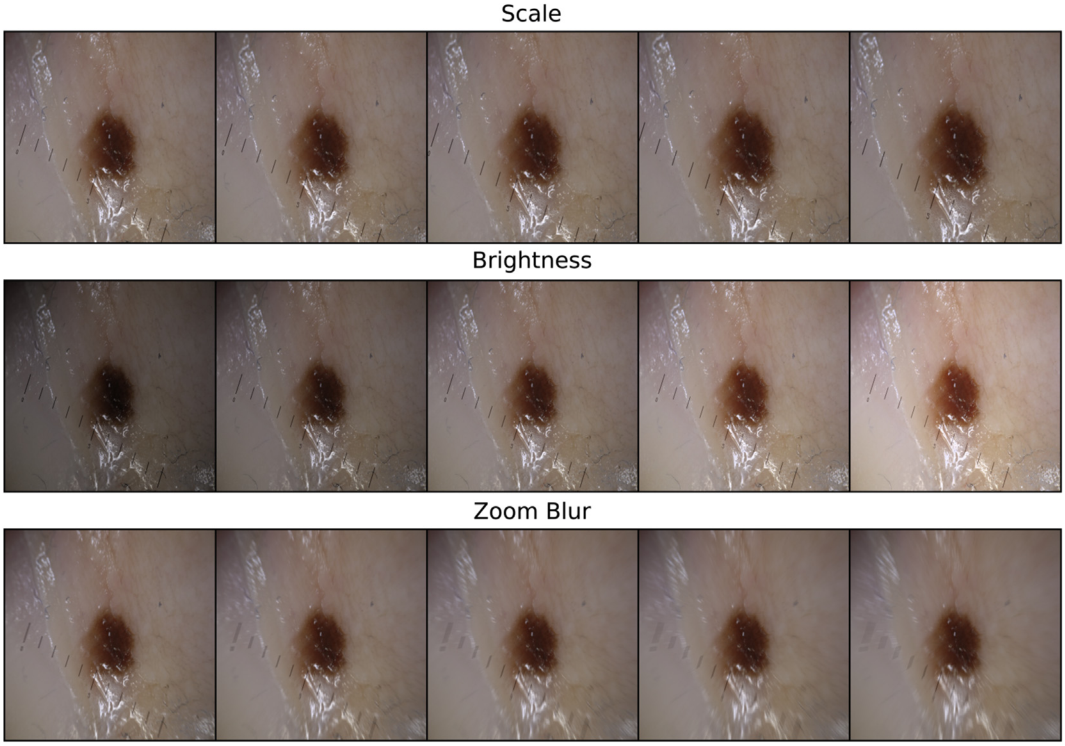

Fig. 2. **Overview of SAM-P perturbations.** Shown here are three perturbation types applied to an exemplary dermoscopic image. We only show a selection of frames for each perturbation sequence (frames 5, 10, 15, 20 and 25). SAM-P, SAM-perturbed.

$$BCE_c^f = \left(\sum_{s=1}^{5} BE_{s,c}^f\right) \Big/ \left(\sum_{s=1}^{5} BE_{s,c}^{AlexNet}\right)$$

We compute BCE_c^f for each corruption type (i.e. $BCE_{Gaussian\ Noise}^f$, $BCE_{Shot\ Noise}^f$, etc.) and finally average across all corruption types to obtain the mean balanced corruption error mBCE.

In addition to the mBCE, we also introduce the relative mBCE which measures how gracefully a classifier degrades in the presence of corruptions. For example, there could be a classifier A where the difference between the performance on clean images (BE) and corrupted images (mBCE) is relatively small, showing that the classifier handles corruptions well. Another classifier B could have a large gap between these two metrics, indicating that the classifier does not handle corruptions well. However, if classifier B has a smaller mBCE than classifier A, it seems like it is better equipped to handle corruption even though it degrades more when faced with corruptions. To account for this, the relative BCE is computed as

$$relative\ BCE_c^f = \left(\sum_{s=1}^{5} BE_{s,c}^f - BE_{clean}^f\right) \Big/ \left(\sum_{s=1}^{5} BE_{s,c}^{AlexNet} - BE_{clean}^{AlexNet}\right)$$

and averaged across each corruption type to obtain the relative mBCE.

SAM-P: As every image in SAM-P, which contains m *unmodified* images, is perturbed n times using perturbation p, we obtain m perturbation sequences given as collection $S = \{(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})\}_{i=1}^{m}$ for any given perturbation. A model's flip probability FP_p^f on a collection S of perturbation type p is therefore

$$FP_p^f = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=2}^{n} 1_a\left(f\left(x_j^{(i)}\right) \neq f\left(x_{j-1}^{(i)}\right)\right)$$

For noise perturbation sequences where frames are not temporally related, we do not compute the flip probability with respect to the previous frame, but always with respect to the first unmodified frame. We therefore modify the indicator function to $1_a(f(x_j^{(i)}) \neq f(x_1^{(i)}))$ for all noise perturbation sequences. Similarly to SAM-C, we adjust FP_p^f by dividing with AlexNet's flip probability to obtain the flip rate as $FR_p^f = FP_p^f / FP_p^{AlexNet}$. Finally, we compute FR_p^f for each perturbation type (i.e. $FR_{Gaussian\ Noise}^f$, $FR_{Shot\ Noise}^f$, etc.) and take the average to obtain the mean flip rate mFR.

3. Results

Looking at Fig. 3, we can see that classifier performance on clean images was worst for AlexNet and best for VGG16+BN and ResNet50. DenseNet121 showed the best corruption robustness as both its mBCE and relative mBCE were lowest. Although VGG16+BN and ResNet50 had a lower mBCE than AlexNet, their relative mBCE was substantially higher. With regard to perturbation robustness, AlexNet showed better stability (i.e. lower mFR) than VGG16+BN, ResNet50 or DenseNet121.

To allow for a better understanding of the actual performance effect that is caused by corruptions and perturbations, Supplementary Table 1 provides the unadjusted mBCE and mFR values.

4. Discussion

4.1. General implications

OOD testing is increasingly demanded and used by the machine learning community. Yet, skin cancer studies conducted within the last 3–4 years often lack classifier evaluation on external data [2,5,14,39,40,44]. Therefore, we have created three new data sets—SAM, SAM-C and SAM-P—to establish a robustness benchmark which can be used to evaluate skin lesion classifiers in a true OOD setting.

The SAM data set provides unmodified dermoscopic images and thereby further enriches the already existing collection of dermoscopic data sets [25–30]. This allows classifiers to be tested on images from another clinic likely providing different imaging modalities and patient population. SAM-C and SAM-P enable researchers to test classifiers against corruptions (i.e. low-quality inputs) and perturbations (i.e. small image changes) which may occur in a clinical setting.

By releasing the images without their ground truth labels, we discourage researchers from incorporating parts of the benchmark in their training data so that the OOD nature of this benchmark is preserved. Test results should therefore be sent to us for assessment (see Section 4.4).

4.2. Corruption robustness

In our experiments, we observe that an increase in balanced accuracy on SAM correlates with a lower mBCE. Therefore, architectural improvements seem to improve robustness when taking mBCE as the measure. This trend was also observed in ImageNet classifiers [32] and could suggest that architectures have become better at generalising to corrupted images through learning more invariant decision rules. However, this appears to be only partially true. Looking at the relative mBCE, which measures how gracefully a classifier's balanced accuracy declines in the presence of corruptions, we can see that VGG16+BN and ResNet50 actually perform worse than AlexNet. As both classifiers had better balanced accuracies than AlexNet, their improved corruption robustness (i.e. mBCE) is more likely due to accuracy improvements and not due to these classifiers learning more stable representations. This might only be the case for DenseNet121, which had a lower relative mBCE than AlexNet which was similar to observations made in the study by Hendrycks and Dietterich [32]. This demonstrates that accuracy improvements can mask corruption robustness issues to a certain extent and that the evaluated classifiers, despite seemingly good performance on clean images, are heavily influenced by corruptions.

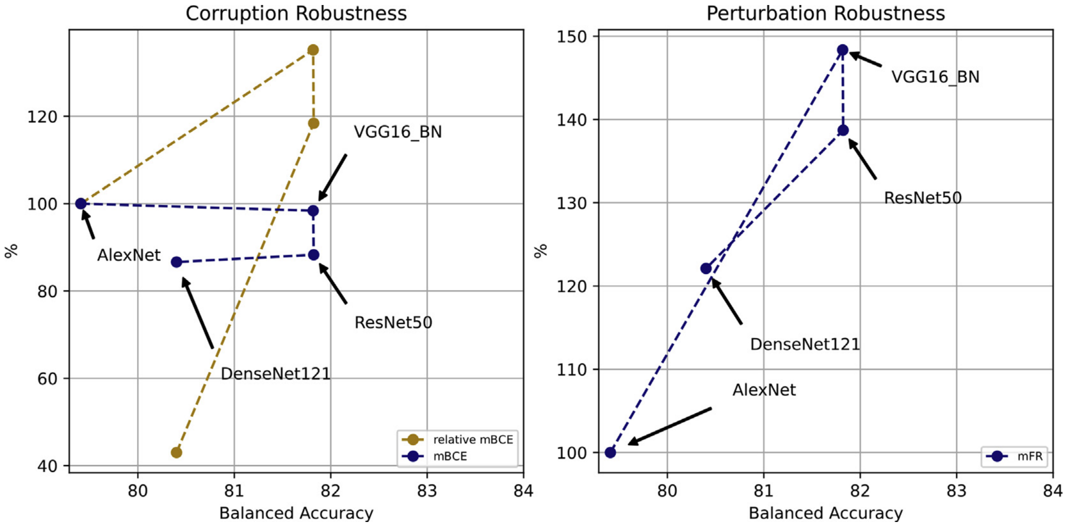

Fig. 3. **Architecture corruption and perturbation robustness on SAM-C and SAM-P, respectively.** Robustness is measured against model performance on unmodified images (SAM) using balanced accuracy. For corruption robustness (left), we plot both the mBCE and the relative mBCE. For perturbation robustness (right), we show the mFR. For mBCE, relative mBCE and mFR lower values are better. SAM, Skin Archive Munich; SAM-C, SAM-corrupted; SAM-P, SAM-perturbed; mBCE, mean balanced corruption error; mFR, mean flip rate.

4.3. Perturbation robustness

SAM-P's transformations were less severe than those used in SAM-C. Nonetheless, classifiers were not consistent in their predictions across perturbation series and showed considerable fluctuations as expected [22,32]. ResNet50, for example, had a 6.7% probability of changing its prediction between adjacent frames, which was quite similar to the other classifiers (see Supplementary Table 1). Looking at Fig. 3, it might be surprising that AlexNet had the lowest mFR since Hendrycks and Dietterich [32] report an opposite trend with ImageNet classifiers. We believe this could be due to the nature of the selected metric which has some downsides. First, a classifier that is heavily biased towards one of the two classes will have a low flip rate and consequently an optimal mFR close to zero. Second, our experiments only have two classes as opposed to 1000 classes for ImageNet. Therefore, a classifier which is just guessing will have a higher chance of predicting the same class a couple of times in a row (i.e. lower flip rate) if there are only two instead of 1000 classes. Both characteristics could affect a classifier's mFR positively. The classifier's accuracy on clean images, in contrast, would presumably suffer in such scenarios. Thus, it is important to consider mFR and performance on clean images in combination.

4.4. How to use this benchmark

A model should be trained on images from any dermoscopic data set(s), excluding images from this benchmark. Training should be carried out without any of the data augmentations that were used to create SAM-C and SAM-P (i.e. no rotations, tilts, Gaussian noise, etc.). Any augmentation overlap should be explicitly stated. The final model should be evaluated as described on GitHub, where we also describe how this benchmark can be obtained (https://github.com/DBO-DKFZ/SCP-Robustness-Benchmark).

5. Conclusions

Using a proprietary data set of 319 dermoscopic melanoma (n = 194) and nevus (n = 125) images, we created three new data sets—SAM, SAM-C and SAM-P—to establish a robustness benchmark which can be used to evaluate skin lesion classifiers in an OOD setting. With SAM-C and SAM-P, classifiers can be tested specifically against image transformations that may occur in a real-life setting. By making the images available without ground truth labels while offering to provide evaluations, we aim to maintain the OOD nature of this benchmark. Altogether, this benchmark makes OOD testing feasible and hopefully standard practice in the future, thereby preparing classifiers for a successful transition to clinical practice.

Role of the funding source

This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Skin Classification Project 2; grant holder: T.J.B., German Cancer Research Center, Heidelberg, Germany). The sponsor had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript and decision to submit the manuscript for publication.

Conflict of interest statement

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: J.S.U. is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, Leo Pharma, Merck Sharp and Dohme, Novartis, Pierre Fabre and Roche, outside the submitted work. F.M. has received travel support or/and speaker's fees or/and advisor's honoraria from Novartis, Roche, BMS, MSD and Pierre Fabre and research funding from Novartis and Roche. S.H. reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Novartis, Roche, BMS, Amgen and MSD outside the submitted work. A.H. reports clinical trial support, speaker's honoraria or consultancy fees from the following companies: Amgen, BMS, Merck Serono, MSD, Novartis, OncoSec, Philogen, Pierre Fabre, Provectus, Regeneron, Roche, OncoSec, Sanofi Genzyme and Sun Pharma, outside the submitted work. W.S. reports grants from medi GmbH Bayreuth, personal fees from Janssen, grants and personal fees from Novartis, personal fees from Lilly, personal fees from UCB, personal fees from Almirall, personal fees from Leo Pharma and personal fees from Sanofi Genzyme, outside the submitted work. B.S. reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Incyte, Novartis, Roche, BMS and MSD, research funding from BMS, Pierre Fabre Pharmaceuticals and MSD and travel support from Novartis, Roche, BMS, Pierre Fabre Pharmaceuticals and Amgen, outside the submitted work. M.G. has received speaker's honoraria and/or has served as a consultant and/or member of advisory boards for Almirall, argenx, Biotest, Eli Lilly, Janssen Cilag, Leo Pharma, Novartis and UCB, outside the submitted work. T.J.B. reports owning a company that develops mobile apps including the teledermatology services AppDoc (https://online-hautarzt.net) and Intimarzt (https://intimarzt.de): Smart Health Heidelberg GmbH, Handschuhsheimer Landstr. 9/1, 69120

Heidelberg, https://smarthealth.de. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2021.06.047.

Author contribution section

Roman C. Maron: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Validation, Visualisation. **Justin G. Schlager:** Resources, Writing - Review & Editing. **Sarah Haggenmüller:** Validation, Writing – original draft, Visualisation **Christof von Kalle:** Resources, Writing - Review & Editing. **Jochen S. Utikal:** Resources, Writing - Review & Editing. **Friedegund Meier:** Resources, Writing - Review & Editing. **Frank F. Gellrich:** Resources, Writing - Review & Editing. **Sarah Hobelsberger:** Resources, Writing - Review & Editing. **Axel Hauschild:** Resources, Writing - Review & Editing. **Lars French:** Resources, Writing - Review & Editing. **Lucie Heinzerling:** Resources, Writing - Review & Editing. **Max Schlaak:** Resources, Writing - Review & Editing. **Kamran Ghoreschi:** Resources, Writing - Review & Editing. **Franz J. Hilke:** Resources, Writing - Review & Editing. **Gabriela Poch:** Resources, Writing - Review & Editing. **Markus V. Heppt:** Resources, Writing - Review & Editing. **Carola Berking:** Resources, Writing - Review & Editing. **Sebastian Haferkamp:** Resources, Writing - Review & Editing. **Wiebke Sondermann:** Resources, Writing - Review & Editing. **Dirk Schadendorf:** Resources, Writing - Review & Editing. **Bastian Schilling:** Resources, Writing - Review & Editing. **Matthias Goebeler:** Resources, Writing - Review & Editing. **Eva Krieghoff-Henning:** Conceptualisation, Writing - Original Draft, Project administration. **Achim Hekler:** Conceptualisation, Methodology, Writing - Review & Editing, Project administration. **Stefan Fröhling:** Resources, Writing - Review & Editing. **Daniel B. Lipka:** Resources, Writing - Review & Editing. **Jakob N. Kather:** Resources, Writing - Review & Editing. **Titus J. Brinker:** Conceptualisation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

References

[1] Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. NPJ Digit Med 2021;4:5.

[2] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

[3] Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for Benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018;138:1529–38.

[4] Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, 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. Lancet Oncol 2019;20:938–47.

[5] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Canc 2019;113:47–54.

[6] Maron RC, Weichenthal M, Utikal JS, Hekler A, Berking C, Hauschild A, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Canc 2019;119:57–65.

[7] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Canc 2019;111:148–54.

[8] Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Canc 2019;120:114–21.

[9] Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. J Invest Dermatol 2020;140:1753–61.

[10] Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nat Med 2020;26:1229–34.

[11] Maron RC, Utikal JS, Hekler A, Hauschild A, Sattler E, Sondermann W, et al. Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma image classification: web-based survey study. J Med Internet Res 2020;22:e18091.

[12] Phillips M, Marsden H, Jaffe W, Matin RN, Wali GN, Greenhalgh J, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. JAMA Netw Open 2019;2:e1913436.

[13] MacLellan AN, Price EL, Publicover-Brouwer P, Matheson K, Ly TY, Pasternak S, et al. The use of non-invasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. J Am Acad Dermatol 2020;85:353–9. https://doi.org/10.1016/j.jaad.2020.04.019.

[14] Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29:1836–42.

[15] Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 2019;10:1096.

[16] Azulay A, Weiss Y. Why do deep convolutional networks generalize so poorly to small image transformations? arXiv [csCV]; 2018.

[17] Zhang R. Making convolutional networks shift-invariant again. arXiv [csCV]; 2019.

[18] Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D. Natural adversarial examples. arXiv [csLG]; 2019.

[19] Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A. Adversarial examples are not bugs, they are features. arXiv [statML]; 2019.

[20] Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Mach Intel 2020;2:665–73.

[21] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018;15:e1002683.

[22] Maron RC, Haggenmüller S, von Kalle C, Utikal JS, Meier F, Gellrich FF, et al. Robustness of convolutional neural networks in recognition of pigmented skin lesions. Eur J Canc 2021;145: 81–91.

[23] Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol 2019;155:1135–41.

[24] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287–9.

[25] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5:180161.

[26] Gutman D, Codella NCF, Celebi E, Helba B, Marchetti M, Mishra N, et al. Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv [csCV]; 2016.

[27] Combalia M, Codella NCF, Rotemberg V, Helba B, Vilaplana V, Reiter O, et al. BCN20000: dermoscopic lesions in the wild. arXiv [eessIV]; 2019.

[28] Mendonca T, Ferreira PM, Marques JS, Marcal ARS, Rozeira J. PH^2 - a dermoscopic image database for research and benchmarking. In: 2013 35th annual international conference of the IEEE engineering in medicine and biology society. EMBC); 2013. p. 5437–40.

[29] Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. 7-Point checklist and skin lesion classification using multi-task multimodal neural nets. IEEE J Biomed Health Inform 2018;23: 538–46. https://doi.org/10.1109/JBHI.2018.2824327.

[30] de Faria SMM, Henrique M, Filipe JN, Pereira PMM, Tavora LMN, Assuncao PAA, et al. Light field image dataset of skin lesions. Conf Proc IEEE Eng Med Biol Soc 2019;2019: 3905–8.

[31] Deng J, Dong W, Socher R, Li L, Li Kai, Fei-Fei Li. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition; 2009. p. 248–55.

[32] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. arXiv [csLG]; 2019.

[33] Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, Dorundo E, et al. The many faces of robustness: a critical analysis of out-of-distribution generalization. arXiv [csCV]; 2020.

[34] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012;25:1097–105.

[35] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv [csCV]; 2014.

[36] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.

[37] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–8.

[38] Howard J, Gugger S. Fastai: a layered API for deep learning. Information 2020;11:108.

[39] Brinker TJ, Hekler A, Enk AH, von Kalle C. Enhanced classifier training to improve precision of a convolutional neural network to identify images of skin lesions. PLoS One 2019;14: e0218713.

[40] Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78. 270–277.e1.

[41] Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer 2019; 119:11–7.

[42] Hekler A, Utikal JS, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur J Cancer 2019;115:79–83.

[43] Brinker TJ, Schmitt M, Krieghoff-Henning E, Barnhill R, Beltraminelli H, Braun SA, et al. Diagnostic performance of artificial intelligence for histologic melanoma recognition compared to 18 international expert pathologists. J Am Acad Dermatol 2021.

[44] Haggenmüller S, Maron RC, Hekler A, Utikal JS, Barata C, Barnhill RL, et al. Skin Cancer Classification via Convolutional Neural Networks: systematic Review of Studies Involving Human Experts. Eur J Cancer 2021. In press.

[45] Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer 2019;111:30–7.

[46] Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019; 118:91–6.