European Journal of Cancer 145 (2021) 81–91

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Robustness of convolutional neural networks in recognition of pigmented skin lesions

Roman C. Maron [a], Sarah Haggenmüller [a], Christof von Kalle [b], Jochen S. Utikal [c,d], Friedegund Meier [e], Frank F. Gellrich [e], Axel Hauschild [f], Lars E. French [g,p], Max Schlaak [g], Kamran Ghoreschi [h], Heinz Kutzner [i], Markus V. Heppt [j], Sebastian Haferkamp [k], Wiebke Sondermann [l], Dirk Schadendorf [l], Bastian Schilling [m], Achim Hekler [a], Eva Krieghoff-Henning [a], Jakob N. Kather [n], Stefan Fröhling [o], Daniel B. Lipka [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 Clinical-Translational Sciences, Charité University Medicine and Berlin Institute of Health (BIH), Berlin, Germany*
[c] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[d] *Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[e] *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, Germany*
[f] *Department of Dermatology, University Hospital (UKSH), Kiel, Germany*
[g] *Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany*
[h] *Department of Dermatology, Venereology and Allergology, Charité – Universitätsmedizin Berlin, Berlin, Germany*
[i] *Dermatopathology Laboratory, Friedrichshafen, Germany*
[j] *Department of Dermatology, University Hospital Erlangen, Erlangen, Germany*
[k] *Department of Dermatology, University Hospital Regensburg, Regensburg, Germany*
[l] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[m] *Department of Dermatology, University Hospital Würzburg, Würzburg, Germany*
[n] *Division of Translational Medical Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[o] *National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[p] *Dr. Phillip Frost Department of Dermatology and Cutaneous Surgery, University of Miami, Miller School of Medicine, Miami, FL, USA*

Received 5 October 2020; received in revised form 6 November 2020; accepted 15 November 2020
Available online 7 January 2021

* *Corresponding author*: Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, Heidelberg, 69120, Germany.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

https://doi.org/10.1016/j.ejca.2020.11.020

KEYWORDS
Artificial intelligence;
Machine learning;
Deep learning;
Neural networks;
Dermatology;
Skin neoplasms;
Melanoma;
Nevus

Abstract ***Background:*** A basic requirement for artificial intelligence (AI)–based image analysis systems, which are to be integrated into clinical practice, is a high robustness. Minor changes in how those images are acquired, for example, during routine skin cancer screening, should not change the diagnosis of such assistance systems.
Objective: To quantify to what extent minor image perturbations affect the convolutional neural network (CNN)–mediated skin lesion classification and to evaluate three possible solutions for this problem (additional data augmentation, test-time augmentation, anti-aliasing).
Methods: We trained three commonly used CNN architectures to differentiate between dermoscopic melanoma and nevus images. Subsequently, their performance and susceptibility to minor changes ('brittleness') was tested on two distinct test sets with multiple images per lesion. For the first set, image changes, such as rotations or zooms, were generated artificially. The second set contained natural changes that stemmed from multiple photographs taken of the same lesions.
Results: All architectures exhibited brittleness on the artificial and natural test set. The three reviewed methods were able to decrease brittleness to varying degrees while still maintaining performance. The observed improvement was greater for the artificial than for the natural test set, where enhancements were minor.
Conclusions: Minor image changes, relatively inconspicuous for humans, can have an effect on the robustness of CNNs differentiating skin lesions. By the methods tested here, this effect can be reduced, but not fully eliminated. Thus, further research to sustain the performance of AI classifiers is needed to facilitate the translation of such systems into the clinic.

1. Introduction

Artificial intelligence (AI)–based image classification by convolutional neural networks (CNNs) has the potential to assist clinicians with diagnostic tasks that are based on the visual inspection of potentially malignant lesions. In experimental settings, CNNs have achieved performances in medical image classification tasks that were on par or even exceeded the results obtained by human experts [1–4]. In particular, CNNs have shown very promising results in macroscopic and microscopic skin lesion classification, both individually [5–12,44,45] and as assistance systems for dermatologists [13–16]. And while such systems are as of yet, mostly unable to predict malignant oncologic transformations due to a lack of prospective training data [47], they are already used in practice. In fact, CNN-based systems have begun to enter clinical dermatological practice as skin cancer screening tools, for example, as a market-approved computer-aided diagnostic (CAD) system [17], which has demonstrated superior performance to more conventional CAD systems [18].

While CNN-based image analysis has advantages over human observation with respect to objective and quantitative feature extraction, an obvious drawback is that in contrast to human experts, CNNs have difficulty distinguishing biologically significant features from insignificant features and artifacts. Depending on the data set that is used for CNN training, spurious and unwanted correlations within the training set can be picked up and hamper generalization [19,20,46]. Moreover, deceptively created input images specifically designed to fool a CNN (adversarial attacks) have been shown to pose a real threat [21]. Both shortcomings also apply to CNNs in the field of dermatology [22–25].

Another observed shortcoming is the brittleness of modern CNNs in image analysis. Brittleness in this context refers to the phenomenon that small changes in the input image, such as scaling or rotation, can have a large effect on the classification of the CNN. It is therefore different to adversarial attacks, as image changes are not designed to deceive the CNN, but reflect fluctuations in image acquisition occurring in daily clinical routine. The resulting vulnerability of AI-based tools contradicts the assumption that CNNs are invariant to small transformations and is reported in the machine learning community [24–28]. As this lack of robustness and reliability may have a detrimental effect in a clinical setting, it needs to be overcome to facilitate the successful translation of AI-based diagnostic tools into routine clinical care.

In this study, we investigate the brittleness of three commonly used CNN architectures, which could serve as backends of CNN-based diagnostic systems, by testing them on images that have undergone transformations, which model variations that may occur when dermatologists photograph suspicious skin lesions. Moreover, we investigate three possible techniques (data augmentation, test time augmentation, anti-aliased networks) regarding

their effectiveness in solving the problem of CNN brittleness.

2. Materials and methods

2.1. Study design

We trained three commonly used CNN architectures (ResNet50, DenseNet121, VGG16) to distinguish between dermoscopic nevus and melanoma images. To establish the models' susceptibility to image changes, each classifier was evaluated on a test set containing unmodified, original images and several additional sets containing duplicated images that were digitally modified. Transformations were chosen to mimic events, which might occur in a clinical setting. Moreover, the magnitude of transformations was limited to an extent which would not render the lesion unrecognizable for a physician. Subsequently, a range of pre-existing methods which address AI brittleness were tested to assess if they are indeed effective in reducing brittleness without impairing performance.

As the test set transformations described above were artificial, the models and methods were additionally tested on an independent test set where at least two dermoscopic images with natural changes resulting from differences in real-life image acquisition were available for each lesion.

Ethics approval was waived by the ethics committee of the University of Heidelberg, as images were open source and anonymous.

2.2. Data sets

Dermoscopic images were obtained from the ISIC archive [29], the HAM10000 data set [30], the PH2 data set [31], the SKINL2 data set [32], the BCN20000 data set [33], and PROP, a proprietary data set. The training set was made up exclusively of ISIC, HAM10000, and BCN20000 images. The artificial test set consisted of a holdout component in ISIC and an external component in PH2 and SKINL2. Similarly, the natural test set consisted of a holdout component in BCN20000 and an external component in PROP. Exact details on training and test set composition are listed in Supplementary Table 1.

The artificial test set was duplicated 11 times. Each of the 11 duplicated sets was modified according to one previously defined transformation type and magnitude. Available types were a change in orientation, zoom, or brightness. In addition, the artificial test set was duplicated six more times, but this time combinations of transformations were applied and the magnitude was increased (see Supplementary Methods and Supplementary Fig. 1).

The additional natural test set contained at least two separately taken dermoscopic images per lesion. Thus, the changes between these images were not produced retrospectively using a computer. As this makes it impossible to define an original test set against which deviations should be measured, all possible image combinations were compiled and evaluated. Because different photographs of the same lesion often looked extremely different, e.g. because of an altered zoom by more than 50%, images for each lesion were manually sorted into similarly looking groups using the four-eyes principle.

2.3. Classifier development

All classifiers, regardless of architecture, were trained using the same training set and protocol. Furthermore, all architectures had the same set of fully connected layers on top of the individual feature extractor, which was made up of fastai's [34] default custom head. Online data augmentation was applied during training, where the type and magnitude of augmentations were adapted from the fastai library, which has sensible preset values. For exact details on the training procedure and used augmentations, see Supplementary Methods.

All work was carried out in Python 3.7.7 using fastai 1.0.61 in combination with torch 1.5.1 [35] and torchvision 0.6.1. Training was carried out on a single NVIDIA GeForce RTX 2080 Ti.

2.4. Methods to reduce brittleness

Three methods were tested for their effectiveness against brittleness. The first approach used a more extreme form of data augmentation during the training stage, where the magnitudes of the applied transformations were increased. The second approach used test-time augmentation during the inference stage. Instead of the model just rating one version of an input image, it rates a collection of slightly modified duplicates and averages the output. In our case, eight modified duplicates were rated, which were transformed using a flip coupled with a zoom into all four image corners. These transformations were set to be deterministic to allow reproducibility. The third approach replaced the original model architecture by an anti-aliased architecture, which reduces anti-aliasing effects in downsampling layers (strided convolutions, max-/average-pooling) [24]. This is achieved by upgrading all downsampling layers to include a low-pass filter. While originally intended to address shift-invariance, a general positive effect on model robustness was observed [24].

2.5. Analysis

To obtain robust performance estimates that encompass the stochastic nature of the training process, each

training and evaluation run was repeated five times. Thus, all calculated metrics are averaged over five runs.

Classifier performance was captured using the area under receiver operating characteristic (AUROC). As the receiver operating curve shows the sensitivity and specificity of a dichotomous outcome for all possible classification thresholds, the area under this curve provides a single summary measure, which captures a classifier's overall performance. The classifiers' susceptibility to change was measured using P(class change) and mean absolute change, two metrics adapted from Ref. [27]. P(class change) represents the probability that the classifier changes its prediction from melanoma to nevus or vice versa, after the input image is transformed. This measure is independent of small confidence fluctuations, which do not have an impact on the classification, e.g. when a model changes its lesion diagnosis from 95% nevus to 85% nevus, this change is ignored by P(class change). Mean absolute change measures by how much on average the model's output probability changes after the input image is transformed. This metric allows us to verify if class changes are mainly a result of lesions being diagnosed divergently when the model was unsure to begin with. For a robust classifier, both metrics should be minimised.

3. Results

3.1. Baseline performance and brittleness

All baseline CNNs achieved an AUROC of approximately 0.9. This was comparable with the AUROCs obtained across the 11 artificially transformed test sets (see Fig. 1, Supplementary Figs. 1 and 2). For ResNet50, the mean absolute change varied from 2.9% ± 0.4% to 11.2% ± 1.2% and resulted in a P(class change) ranging from 3.5% ± 0.9% to 12.2% ± 1.6%. Variations in mean absolute change and P(class change) were slightly lower for DenseNet121, with VGG16 showing the lowest variation out of all three architectures (see Supplementary Figs. 1 and 2).

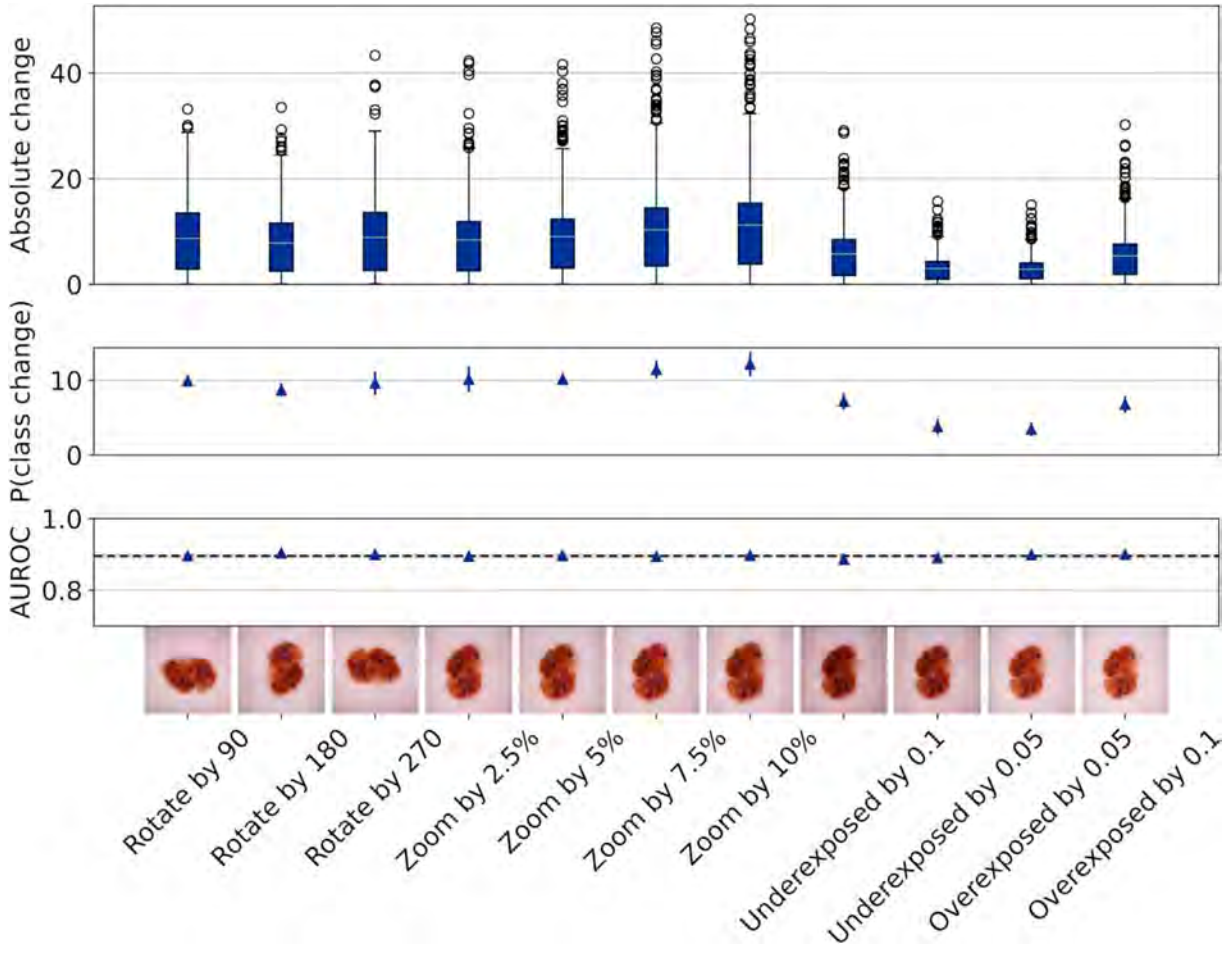

Fig. 1. **Individual performance and brittleness metrics for the baseline ResNet50 model across all artificially transformed test sets.** Top row shows the absolute change distribution over each artificially transformed test set. The grey line within the box plot indicates the mean absolute change. Middle row and bottom row show the mean P(class change) and AUROC, respectively, for each individually and artificially transformed test set. In addition, the AUROC for the unmodified test set is shown as a dashed line. Results for the other architectures were similar (see Supplementary Figs. 1 and 2).

Averaging performance and robustness metrics across all twelve artificial test sets shows that both metrics were always better on the holdout than on the external test set regardless of used architecture (see Supplementary Table 2). Moreover, there was a clear ranking between architectures with VGG16 having the best overall performance and brittleness scores, followed by DenseNet121 and ResNet50.

3.2. Effectiveness of tested methods on artificial transformations

The three tested methods, which were additional data augmentation, test-time augmentation and anti-aliasing, were able to reduce overall brittleness when applied individually and especially when used in combination. This was true for all three architectures to a similar extent and did not result in performance deterioration (see Fig. 2). Depending on the type of transformation that was applied to the test set, the used methods showed varying degrees of effectiveness. Generally, larger improvements were observed for rotations and zooms than for brightness (see Supplementary Fig. 4).

When combining the artificial transformations to act on an image together, brittleness increased even more and performance deteriorated slightly. However, all reviewed methods were still effective in reducing brittleness while upholding performance (see Fig. 3).

Regardless whether the artificial transformations were used individually or in combination, additional data

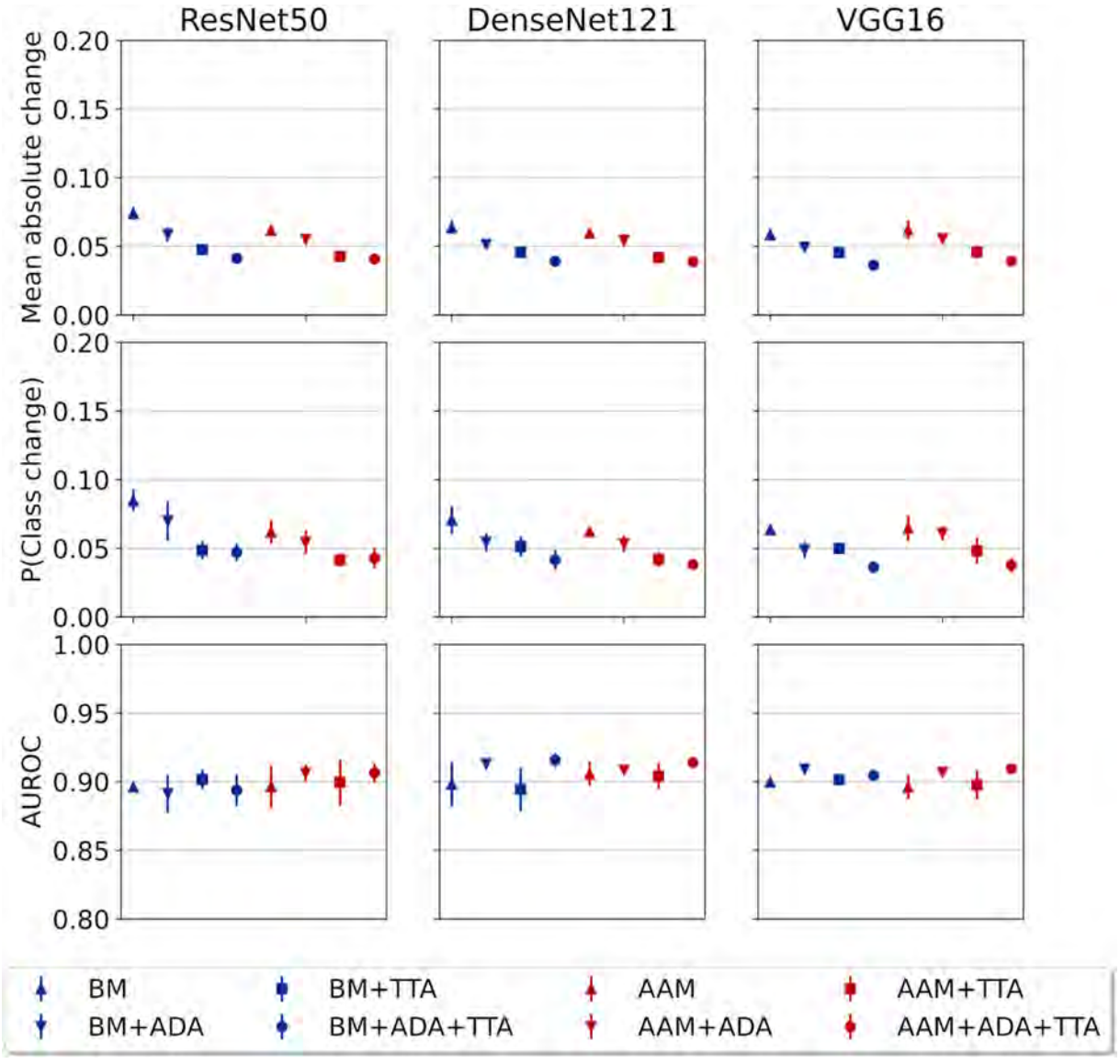

Fig. 2. **Average performance and brittleness metrics across all artificially transformed test sets for the various method combinations using individual transformations.** The three proposed methods, ADA, TTA and AAM were tested individually and in combination. Metrics were established on all individually transformed test sets and averaged. BM: baseline model, ADA: additional data augmentation, TTA: test-time augmentation, AAM: anti-aliased model.

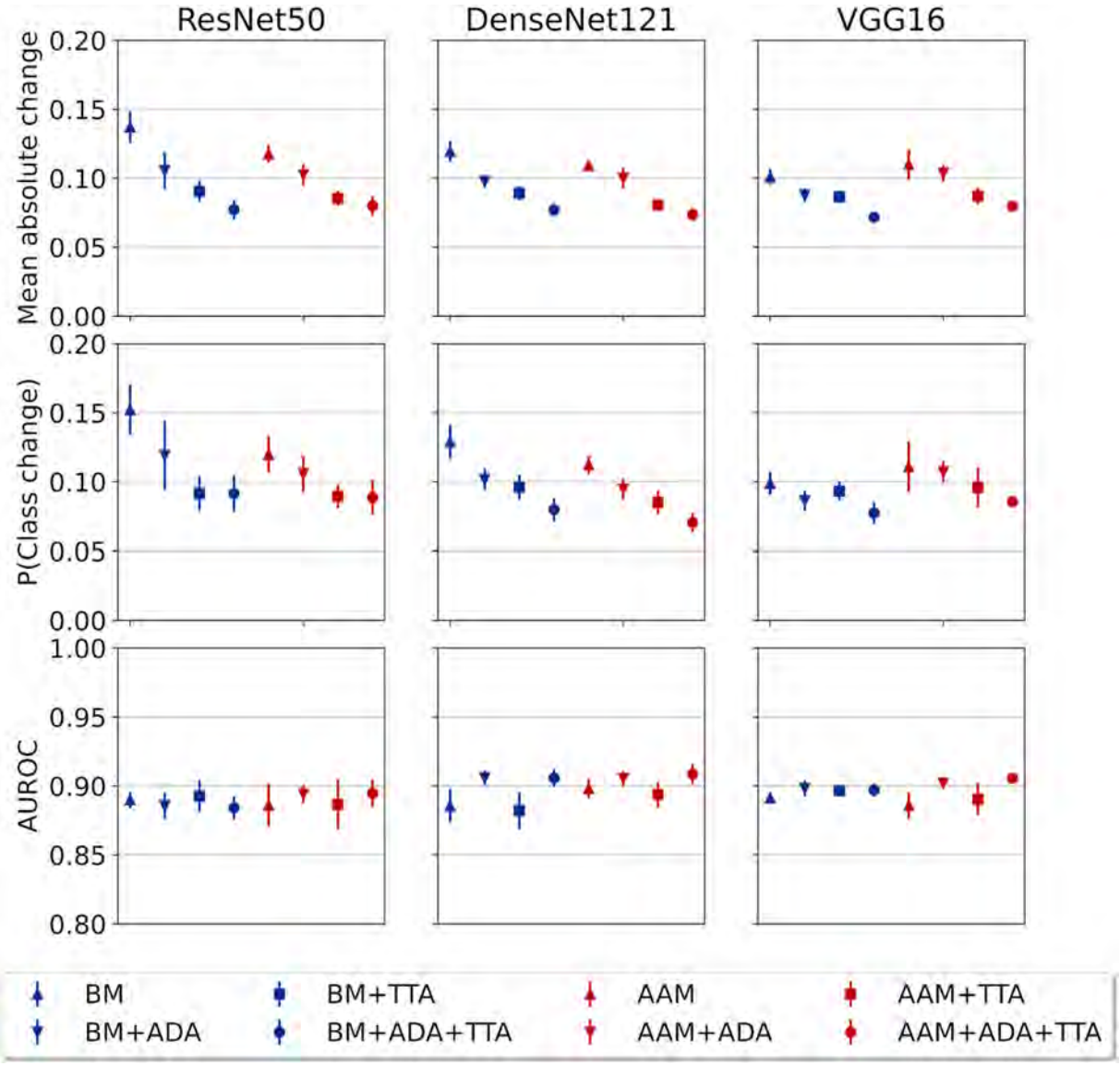

Fig. 3. **Average performance and brittleness metrics across all artificially transformed test sets for the various method combinations using combined transformations.** The three proposed methods, ADA, TTA and AAM were tested individually and in combination. Metrics were established and averaged over all transformed test sets, which were modified using a combination of individual transformations. BM: baseline model, ADA: additional data augmentation, TTA: test-time augmentation, AAM: anti-aliased model.

augmentation and test-time augmentation always showed improvements for brittleness and were most effective when applied in combination. Anti-aliasing worked well for ResNet50 and DenseNet121; however, the anti-aliased VGG16 suffered an increase in brittleness.

3.3. Effectiveness of tested methods on natural transformations

Average performance and brittleness of all three baseline models on the natural test set was in-between that of the artificial test set with individual transformations and the artificial test set with combined transformations. However, effectiveness of the employed methods was far less pronounced on the natural test set than on either of the two artificial test sets (see Fig. 4). Trends were less consistent and while one method showed improvements for a certain architecture, it did not do so for another. For example, ResNet50 experienced slightly worse brittleness with additional data augmentation while DenseNet121 did not. Regardless of architecture, test-time augmentation always improved both performance and brittleness.

4. Discussion

4.1. Practical implications

This study demonstrated brittleness i.e. vulnerability of CNNs toward small input changes for three commonly

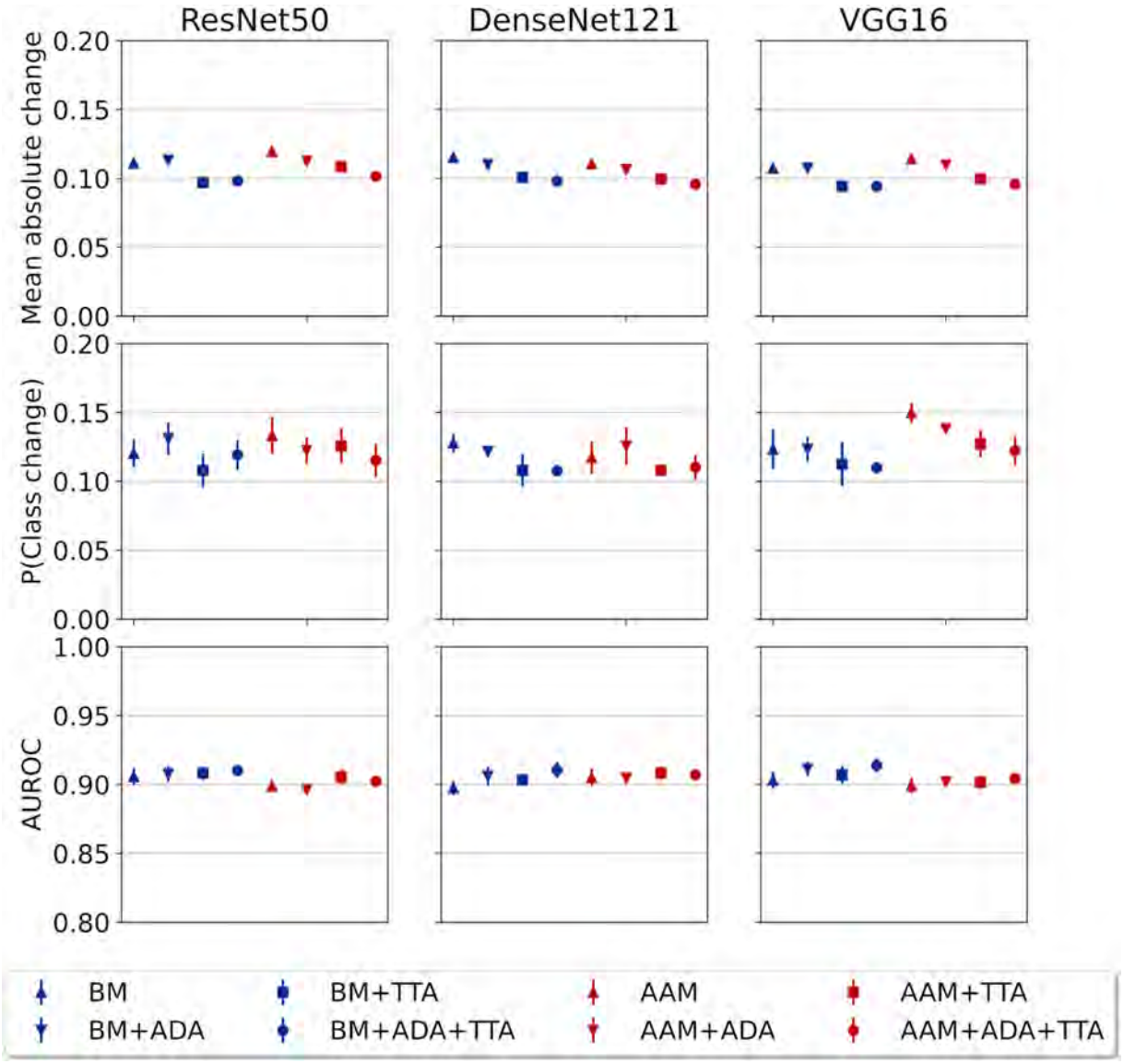

Fig. 4. **Performance and brittleness metrics across the natural transformed test set for the various method combinations.** The three proposed methods, ADA, TTA and AAM, were tested individually and in combination. BM: baseline model, ADA: additional data augmentation, TTA: test-time augmentation, AAM: anti-aliased model.

used CNN architectures (ResNet50, DenseNet121, VGG16). Although this phenomenon has been reported throughout the machine learning community, its potential impact on AI-based assistance systems in the clinic has not received proper attention [24–28]. We reviewed three different methods to reduce brittleness (additional data augmentation, test-time augmentation, anti-aliasing) and found them to be partially effective on artificial image transformations such as rotations, altered brightness or zooms, but less so on natural image transformations resulting from image acquisition differences.

For our models, we chose architectures and training techniques that are commonly used throughout image classification tasks for skin cancer [6,15,36] and other cancer subtypes [2,37–40]. Thus, we believe our baseline models to be suitably representative of existing or future models, which could serve as the backbone of a diagnostic system.

While the change of diagnosis i.e. P(class change) is independent of monotonic confidence fluctuations and intuitive to grasp, we also consider the mean absolute change. In a clinical setting, it is unlikely that an assistance system, which solely presents a plain diagnosis such as melanoma, will be accepted by physicians or patients. Inclusion of the model's confidence level may increase trust in the system as it enables the physician to judge the weight he/she should attribute to the model's classification. Low-confidence decisions by the system would therefore be less likely to influence the physician's

management decision to begin with. In such a setting, brittleness would be partially compensated as the observed confidence changes would often only alter the CNN's classification if its confidence was low to begin with. If, however, high-confidence classifications show these fluctuations, the range of confidences for similar images can be highly disconcerting to the physician.

The techniques we evaluated to reduce brittleness, namely additional data augmentation, test-time augmentation and anti-aliasing, substantially reduced this phenomenon in an artificial setting, but even when used in combination did not completely eliminate it. Depending on the architecture, some methods worked better than others; for example, anti-aliasing did not reduce brittleness for VGG16. When all three methods were used in combination, brittleness and performance always improved in comparison with the baseline model.

The observed improvement was much more limited on naturally transformed images. Even when combinations were applied, improvements were minor or non-existent. Fig. 5 shows a selection of natural image pairs where our models, regardless of the applied method, always came to a divergent diagnosis on an image pair of the same lesion, even though some of the paired images appear almost identical. Thus, it may hardly be possible for a physician to determine how to photograph a lesion 'correctly', which they intend to diagnose with a CNN-based lesion classification system. Such problems limit the applicability of the technology in the clinic and therefore have to be solved.

Against this background, we would like to inform physicians to not consider CNN-based systems as error free and be aware of such limitations. We also want to encourage deep learning practitioners to actively minimise brittleness on a case-by-case basis in the same way performance is optimised. The reported improvements could be further enhanced through method-specific optimisations, alternative techniques for robustness [41,42] or an ensemble-approach, which showed even better improvements than model-specific techniques (see Supplementary Table 3). Finally, future work should also investigate alternatives, which do not solely focus on the training/inference procedure or on architectural modifications but rather on other architectures such as Capsule Neural Networks [43] which could be better suited to handle small affine transformations.

4.2. Limitations

The artificial image changes were designed in such a way as to be relatively inconspicuous to a human observer. The inconspicuousness was determined using the four-eye principle and is therefore subjective. But even if images changes are not deemed as inconspicuous, such transformations are still likely to arise in a clinical setting and therefore any CNN-based system should be invariant against such changes.

The natural test set contained multiple photographs per lesion, where some looked extremely distinct, to the point where there was no overlap between images. Thus,

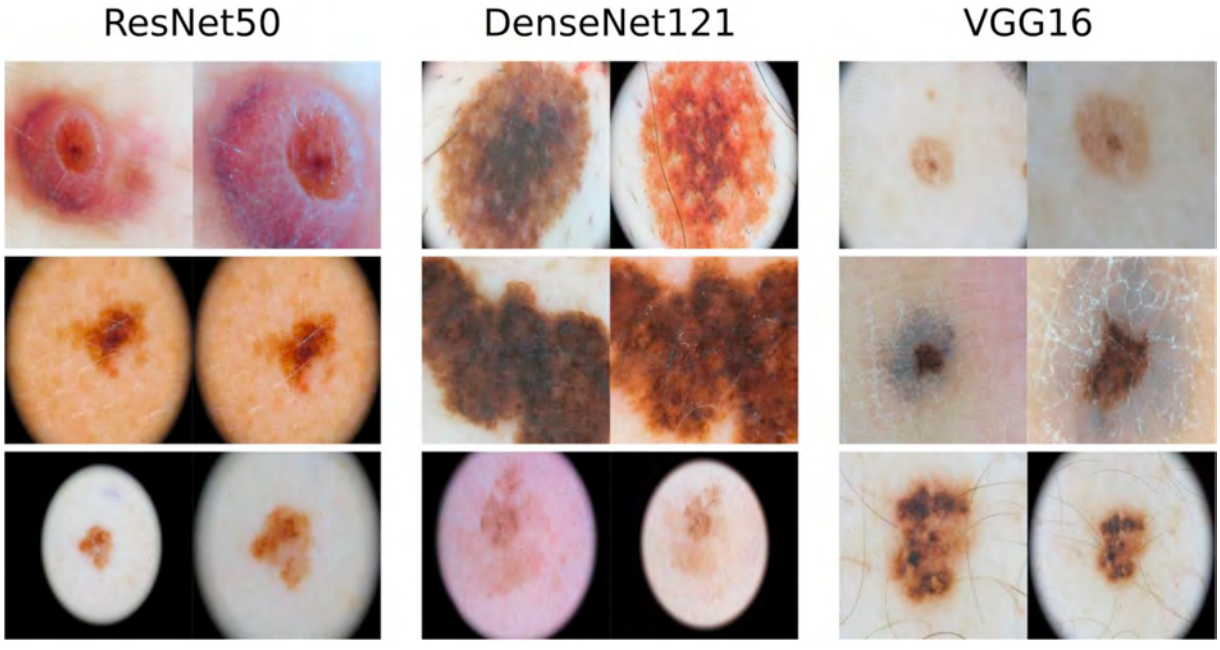

Fig. 5. **Natural image pairs for selected lesions where models disagreed constantly.** Each lesion was photographed twice and rated by all possible combinations of proposed methods (i.e. BM, BM + ADA, BM + TTA, AAM, etc.). Regardless of the applied method, none of the selected image pairs received the same diagnosis. BM: baseline model; ADA: additional data augmentation, TTA: test-time augmentation, AAM: anti-aliased model.

suitably similar image pairs for each lesion were manually chosen using the four-eye principle. As this was largely subjective, the reported results for the natural test set could change depending on how the images are sorted.

5. Conclusions

Minor image changes, relatively inconspicuous for humans, can have an effect on the confidence and diagnosis of CNNs differentiating skin lesions. Using the methods tested here, this effect was reduced but not fully eliminated. Therefore, we would like to remind deep learning practitioners and physicians in dermatology but also in medicine in general, that brittleness needs to be explicitly targeted and overcome to facilitate translation from bench-to-bedside.

Role of the funding source

This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Skin Classification Project; grant holder: Titus J. Brinker, 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

Sebastian H. reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Novartis, Roche, BMS, Amgen and MSD outside the submitted work. Axel 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. BS 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. JSU is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, LeoPharma, Merck Sharp and Dohme, Novartis, Pierre Fabre, Roche, outside the submitted work. WS received travel expenses for attending meetings and/or (speaker) honoraria from Abbvie, Almirall, Bristol-Myers Squibb, Celgene, Janssen, LEO Pharma, Lilly, MSD, Novartis, Pfizer, Roche, Sanofi Genzyme and UCB outside the submitted work. FM has received travel support or/and speaker's fees or/and advisor's honoraria by Novartis, Roche, BMS, MSD and Pierre Fabre and research funding from Novartis and Roche. TJB reports owning a company that develops mobile apps (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.

CRediT authorship contribution statement

Roman C. Maron: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft. **Sarah Haggenmüller:** Conceptualization, Methodology, Writing - original draft, Visualization. **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. **Axel Hauschild:** Resources, Writing - review & editing. **Lars E. French:** Resources, Writing - review & editing. **Max Schlaak:** Resources, Writing - review & editing. **Kamran Ghoreschi:** Resources, Writing - review & editing. **Heinz Kutzner:** Resources, Writing - review & editing. **Markus V. Heppt:** 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. **Achim Hekler:** Conceptualization, Methodology, Writing - review & editing. **Eva Krieghoff-Henning:** Conceptualization, Writing - original draft. **Jakob N. Kather:** Resources, Writing - review & editing. **Stefan Fröhling:** Resources, Writing - review & editing. **Daniel B. Lipka:** Resources, Writing - review & editing. **Titus J. Brinker:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2020.11.020.

References

[1] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.

[2] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199–210.

[3] Tseng H-H, Wei L, Cui S, Luo Y, Ten Haken RK, El Naqa I. Machine learning and imaging informatics in oncology. Oncology 2020;98:344–62.
[4] Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, et al. Artificial intelligence-based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Pathol Lab Med 2019;143:859–68.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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 Canc 2019;111:30–7.
[11] 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.
[12] 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 Canc 2019; 119:11–7.
[13] 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.
[14] Han SS, Park I, Chang SE, 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.
[15] 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.
[16] 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.
[17] 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.
[18] Sies K, Winkler JK, Fink C, Bardehle F, Toberer F, Buhl T, et al. Past and present of computer-assisted dermoscopic diagnosis: performance of a conventional image analyser versus a convolutional neural network in a prospective data set of 1,981 skin lesions. Eur J Canc 2020;135:39–46.
[19] 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.
[20] 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.
[21] Heaven D. Why deep-learning AIs are so easy to fool. Nature 2019;574:163–6.
[22] 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. https://doi.org/10.1001/jamadermatol.2019.1735.
[23] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287–9.
[24] Fawzi A, Frossard P. Manitest: are classifiers really invariant? Proc Br Mach Vis Conf 2015 2015:106. https://doi.org/10.5244/c.29.106.
[25] Zhang R. Making convolutional networks shift-invariant again 06. 2019.
[26] Alcorn MA, Li Q, Gong Z, Wang C, Mai L, Ku W-S, et al. Strike (with) a pose: neural networks are easily fooled by strange poses of familiar objects. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR) 2019; 2019. https://doi.org/10.1109/cvpr.2019.00498.
[27] Azulay A, Weiss Y. Why do deep convolutional networks generalize so poorly to small image transformations? arXiv [csCV] 2018.
[28] Engstrom L, Tran B, Tsipras D, Schmidt L, Madry A. Exploring the landscape of spatial robustness. In: Chaudhuri K, Salakhutdinov R, editors. Long beach. vol. 97. California, USA: PMLR; 2019. p. 1802–11.
[29] 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.
[30] 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.
[31] Mendonca T, Ferreira PM, Marques JS, Marcal ARS, Rozeira J. PH2 – 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. https://doi.org/10.1109/embc.2013.6610779.
[32] 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.
[33] Combalia M, Codella NCF, Rotemberg V, Helba B, Vilaplana V, Reiter O, et al. BCN20000: dermoscopic lesions in the wild. arXiv [eessIV] 2019.
[34] Howard J, Gugger S. Fastai: a layered API for deep learning. Information 2020;11:108.
[35] Paszke A, Gross S, Massa F, Lerer A. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 2019;32:8024–35.
[36] Bi L, Kim J, Ahn E, Feng D. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv [csCV] 2017.
[37] Li X, Shen X, Zhou Y, Wang X, Li T-Q. Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet). PloS One 2020;15:e0232127.
[38] Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis C-A, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med 2019;16:e1002730.

[39] Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Can 2020;1:789–99.
[40] Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20:193–201.
[41] Lee J, Won T, Lee TK, Lee H, Gu G, Hong K. Compounding the performance improvements of assembled techniques in a convolutional neural network. arXiv [csCV] 2020.
[42] Lopes RG, Yin D, Poole B, Gilmer J, Cubuk ED. Improving robustness without sacrificing accuracy with patch Gaussian augmentation. arXiv [csLG] 2019.
[43] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural information processing systems. vol. 30. Curran Associates, Inc.; 2017. p. 3856–66.
[44] 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.
[45] 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.
[46] Schmitt M, Maron RC, Hekler A, Stenzinger A, Hauschild A, Weichenthal M, et al. Hidden variables in deep learning digital pathology and their potential to cause batch effects: technical model study. J Med Internet Res (forthcoming) 2021. https://doi.org/10.2196/23436.
[47] Sondermann W, Utikal JS, Enk AH, Schadendorf D, Klode J, Hauschild A, et al. Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: A call for prospective data. Eur J Cancer 2019;119:30–4.