Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge

Marc Combalia, Noel Codella*, Veronica Rotemberg*, Cristina Carrera, Stephen Dusza, David Gutman, Brian Helba, Harald Kittler, Nicholas R Kurtansky, Konstantinos Liopyris, Michael A Marchetti, Sebastian Podlipnik, Susana Puig, Christoph Rinner, Philipp Tschandl, Jochen Weber, Allan Halpern*, Josep Malvehy**

Summary

Background Previous studies of artificial intelligence (AI) applied to dermatology have shown AI to have higher diagnostic classification accuracy than expert dermatologists; however, these studies did not adequately assess clinically realistic scenarios, such as how AI systems behave when presented with images of disease categories that are not included in the training dataset or images drawn from statistical distributions with significant shifts from training distributions. We aimed to simulate these real-world scenarios and evaluate the effects of image source institution, diagnoses outside of the training set, and other image artifacts on classification accuracy, with the goal of informing clinicians and regulatory agencies about safety and real-world accuracy.

Methods We designed a large dermoscopic image classification challenge to quantify the performance of machine learning algorithms for the task of skin cancer classification from dermoscopic images, and how this performance is affected by shifts in statistical distributions of data, disease categories not represented in training datasets, and imaging or lesion artifacts. Factors that might be beneficial to performance, such as clinical metadata and external training data collected by challenge participants, were also evaluated. 25 331 training images collected from two datasets (in Vienna [HAM10000] and Barcelona [BCN20000]) between Jan 1, 2000, and Dec 31, 2018, across eight skin diseases, were provided to challenge participants to design appropriate algorithms. The trained algorithms were then tested for balanced accuracy against the HAM10000 and BCN20000 test datasets and data from countries not included in the training dataset (Turkey, New Zealand, Sweden, and Argentina). Test datasets contained images of all diagnostic categories available in training plus other diagnoses not included in training data (not trained category). We compared the performance of the algorithms against that of 18 dermatologists in a simulated setting that reflected intended clinical use.

Findings 64 teams submitted 129 state-of-the-art algorithm predictions on a test set of 8238 images. The best performing algorithm achieved 58·8% balanced accuracy on the BCN20000 data, which was designed to better reflect realistic clinical scenarios, compared with 82·0% balanced accuracy on HAM10000, which was used in a previously published benchmark. Shifted statistical distributions and disease categories not included in training data contributed to decreases in accuracy. Image artifacts, including hair, pen markings, ulceration, and imaging source institution, decreased accuracy in a complex manner that varied based on the underlying diagnosis. When comparing algorithms to expert dermatologists (2460 ratings on 1269 images), algorithms performed better than experts in most categories, except for actinic keratoses (similar accuracy on average) and images from categories not included in training data (26% correct for experts *vs* 6% correct for algorithms, p<0·0001). For the top 25 submitted algorithms, 47·1% of the images from categories not included in training data were misclassified as malignant diagnoses, which would lead to a substantial number of unnecessary biopsies if current state-of-the-art AI technologies were clinically deployed.

Interpretation We have identified specific deficiencies and safety issues in AI diagnostic systems for skin cancer that should be addressed in future diagnostic evaluation protocols to improve safety and reliability in clinical practice.

Funding Melanoma Research Alliance and La Marató de TV3.

Lancet Digit Health 2022; 4: e330–39

*Contributed equally

Melanoma Unit, Dermatology Department, Hospital Clínic Barcelona, Universitat de Barcelona, CIBER de Enfermedades raras IDIBAPS, Barcelona, Spain (M Combalia MS, C Carrera MD, S Podlipnik MD, Prof S Puig MD, Prof J Malvehy MD)**; Microsoft, Redmond, WA, USA** (N Codella PhD)**; Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA** (V Rotemberg MD, S Dusza DrPH, N R Kurtansky BS, M A Marchetti MD, J Weber, Prof A Halpern MD)**; Emory University School of Medicine, Department of Biomedical Informatics, Atlanta, GA, USA** (D Gutman MD)**; Kitware, Clifton Park, NY, USA** (B Helba BS)**; Department of Dermatology** (H Kittler MD, P Tschandl MD) **and Center for Medical Statistics, Informatics and Intelligent Systems** (C Rinner PhD)**, Medical University of Vienna, Vienna, Austria; University of Athens Medical School, Department of Dermatology-Venereology, Athens, Greece** (K Liopyris MD)

Correspondence to:
Dr Veronica Rotemberg, Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10021, USA
rotembev@mskcc.org

Introduction

Melanoma has the highest mortality rate of all skin cancers, with about 220 000 cases and 37 000 deaths reported annually in the USA and Europe combined.[1] Early detection of melanoma and other skin tumours is the most important predictor for survival.[2,3] Diagnosis of skin cancer requires sufficient expertise and proper equipment for adequate accuracy. For expert dermatologists, the accuracy of melanoma diagnosis is about 71% with naked-eye inspection, and 90% using a dermatoscope, which is a magnifying lens with either liquid emulsion or cross-polarisation filters to eliminate

Research in context

Evidence before this study

We searched arXiv and PubMed Central for articles published in English between Jan 1, 2002, and Feb 28, 2021, using the search terms "melanoma diagnosis" or "melanoma detection". Our search returned more than 60 000 articles, of which 30 were relevant to this topic. The summary estimate of the accuracy of machine learning algorithms reported for melanoma detection had consistently exceeded that of human experts since 2018. We found no study that evaluated algorithm performance across a range of image artifacts and source institutions. Although there were studies that evaluated algorithm performance on untrained diagnostic classes, none systematically evaluated the errors that algorithms are prone to make on untrained images or images with artifacts. We found many studies that were susceptible to biases, such as selection and labelling, and many did not include publicly available data.

Added value of this study

This study provides an analysis of state-of-the-art deep learning algorithms. Using algorithms submitted via the 2019 International Skin Imaging Collaboration Grand Challenge, we assessed the effect on diagnostic accuracy of alternate statistical distributions of data (via different image sources), disease categories not represented in training datasets, imaging or lesion artifacts, and factors that might be beneficial to performance (such as clinical metadata and external training data). Although algorithms continued to outperform expert readers on melanoma detection, shifted statistical distributions and disease categories not included in training data contributed to decreases in algorithm accuracy. For automated methods, around 50% of the images from categories not included in the training data were misclassified as malignant diagnoses, which would lead to a substantial number of unnecessary biopsies if clinically deployed.

Implications of all the available evidence

We have identified specific deficiencies and safety issues in AI dermatological diagnostic systems that should be addressed in future diagnostic evaluation protocols to improve safety and reliability before clinical implementation. These findings advance existing evidence as they highlight the effects of image artifacts, image source institution, and underlying training distributions and diagnostic classes on algorithm performance. This work advocates for future funding and research devoted to accurate benchmarking and predeployment testing that mimics clinical scenarios.

surface reflectance of skin.[4,5] However, there is a global shortage of expert dermatologists: in Spain, there are 3·27 dermatologists per 100 000 citizens, 6·6 in Germany, 0·55 in the UK, and 0·33 in the USA.[6]

Because of this shortage of expertise, efforts have focused on scaling expertise by developing tools for automated assessment. The International Skin Imaging Collaboration (ISIC) Archive has collated the largest public repository of dermoscopic image datasets to support this continued research effort, facilitating 5 years of public challenges to use artificial intelligence (AI) to detect skin cancer.[7–12] Several articles have reported the development of AI systems with diagnostic accuracy superior to expert dermatologists in controlled experiments.[9,13–17]

Although tremendous technical progress has been achieved, there are still important deficiencies that remain to be addressed before clinical deployment. For example, external validation studies with shifted statistical distribution that is more reflective of real-world clinical application have not been performed, even for algorithms that are already available for use in clinical practice.[9,17,18] In addition, current AI systems are unable to communicate what they do not know. For example, when shown an image of a disease not represented in the training data, the system cannot flag it as a category on which it was not trained, and will instead classify it as one of the conditions it was trained to identify.[19,20] Finally, most previous work on this topic has involved studying system performance on only standardised image data or without correlation to performance of dermatologists.[9,21–23]

For the **Covalic platform** see https://github.com/girder/covalic

We aimed to create the largest public dataset in this domain, the BCN20000 dataset, to design a skin cancer recognition challenge that rigorously evaluates the effects to AI performance of statistical imbalances, images from categories not trained (NT), and clinical data of varying quality, and allows us to analyse the effect of these factors on performance. The public challenge approach was chosen to explore the current state-of-the-art algorithms in skin cancer diagnosis through AI. We investigated the accuracy (1) of state-of-the-art classification methods on datasets specifically designed to better reflect clinical realities than previous studies; (2) of algorithms specifically designed to fail safely by flagging not trained categories; (3) and of algorithms as related to real-world clinically unusual features and other imaging artifacts, such as variations in lighting conditions, clinical markings on the skin, or hair occluding visualisation of the lesion. We also tested the algorithms against dermatologists.

Methods

Study design

We designed a large image classification challenge, the ISIC challenge, to quantify the performance of machine learning algorithms for the task of skin cancer classification from dermoscopic images. The challenge was hosted online using the Covalic platform, where challenge participants could upload their algorithm's diagnostic predictions for each image in the test dataset.

Invitations for submissions were solicited from around the world; calls for submissions were sent via email to

ISIC subscribers and the challenge was publicised on social media and at academic conferences. Challenge participants were permitted to form teams, and allowed to submit diagnostic predictions from up to three distinct algorithms. Unlimited submissions were allowed per algorithm, but only the most recent submission was scored. Further details of the challenge can be found online.

We divided the challenge into two tasks: (1) skin cancer classification from dermoscopic images and (2) skin cancer classification from dermoscopic images and metadata.[24] In both tasks, algorithms were tested on their ability to recognise the eight trained categories, as well as whether they were able to fail safely by correctly identifying diagnostic categories on which they were not trained. To improve the reproducibility of successful algorithms, each team in the challenge was required to submit a manuscript detailing the methods used for image classification.[25]

The study protocol was approved by the ethics review boards of the University of Queensland, Memorial Sloan Kettering Cancer Center, the Medical University of Vienna, and the Hospital Clinic of Barcelona. At all contributing institutions, written informed consent for retrospectively collected dermoscopic images was waived by the ethics review due to the deidentified nature of the images.

Datasets

Dermoscopic images of skin lesions were obtained from skin cancer surveillance clinics around the world, with photographs captured between Jan 1, 2000, and Dec 31, 2018. Each image was paired with metadata regarding the age and sex of the patient, the anatomical location of the lesion, and a lesion identifier. Multiple images acquired from different photographic equipment or on different dates were allowed for a given lesion, mimicking true clinical practice. Lesions were partitioned between training and test sets, balanced by source and diagnostic category in the training dataset.

The training dataset contained 25 331 images, which was composed of data from the Medical University of Vienna (HAM10000)[26] and Hospital Clinic Barcelona (BCN20000).[7,27–29] HAM10000 was used as the benchmark for a previous ISIC challenge in 2018.[8,9] All datasets included labels specifying the clinic that data were acquired from, which is henceforth referred to as the source institution.[26]

An independent, unbalanced, validation dataset of 100 randomly selected dermoscopic images captured between Jan 1, 2000, and Dec 31, 2018 from the Medical University of Vienna was available to challenge participants.[9] These images were not included in the training or test datasets and were provided to challenge participants to validate and debug their algorithm submissions, but the validation dataset was not used for evaluation or further assessments.

The test dataset included 8238 images retrospectively collected from the Hospital Clinic Barcelona (BCN) and the Medical University of Vienna (HAM). Images from Turkey, New Zealand, Sweden, and Argentina were also included. Patient images were not individually labelled for ethnicity, skin tone, or nationality.[9] The test dataset contained all diagnostic categories available in training, as well as other diagnoses not included in training data, which were grouped into a single category referred to as NT. Although test data were acquired at centres that also contributed training data, there was no image or lesion overlap between training and testing datasets. Further dataset details and distributions are available in the appendix (p 2).

For the **challenge** see https://challenge.isic-archive.com/landing/2019/

See **Online** for appendix

Diagnostic labels

The training and test datasets contained images of nevi, melanoma, benign keratosis, dermatofibroma, basal cell carcinoma, squamous cell carcinoma including Bowen's disease, vascular lesions, and actinic keratosis. Borderline melanocytic lesions were excluded. Participants were challenged to classify untrained images into a ninth category in the test dataset, labelled NT, which refers to diagnostic classes that were not included in the training data. We generated ground truth diagnostic labels through review of histopathology for all malignant and biopsied lesions and unanimous expert consensus (at least three experts defined as board certified dermatologists from Memorial Sloan Kettering Cancer Center, Medical University of Vienna, or Hospital Clinic Barcelona; VR, CC, MAM, SPo, SPu, JM, PT, and HK), digital monitoring, or confocal microscopy for unbiopsied benign lesions.[7,8] For the BCN dataset, we conducted these reviews. For HAM, we used published data.[26]

Additional labels

In addition to the labels provided as training and testing metadata, geographical characteristics and the source institution were obtained by the researchers of this study for the purposes of this analysis. The source institution represents alternate statistical distributions and photographic acquisition differences.[26] Furthermore, quantified imaging features (such as pigmentation) and lesion artifacts (such as the presence of ulceration, crust, pigmentation, hair, or pen marks) were manually annotated. Paid medical student research fellows at Memorial Sloan Kettering Cancer Center and Hospital Clinic Barcelona used in-house annotation software to annotate the presence or absence of ulceration, crust, pigmentation, hair, or pen using active learning techniques.[9,21,27,30] Pigmentation was defined as a brown pigment in the lesion area, crust was defined as keratinaceous crust or scale over the lesion area, and ulceration was a defect in the epidermal surface (such as an erosion or ulcer). Hair was defined as having vellus or terminal hairs over the lesion of interest, and pen markings could be anywhere in the image.

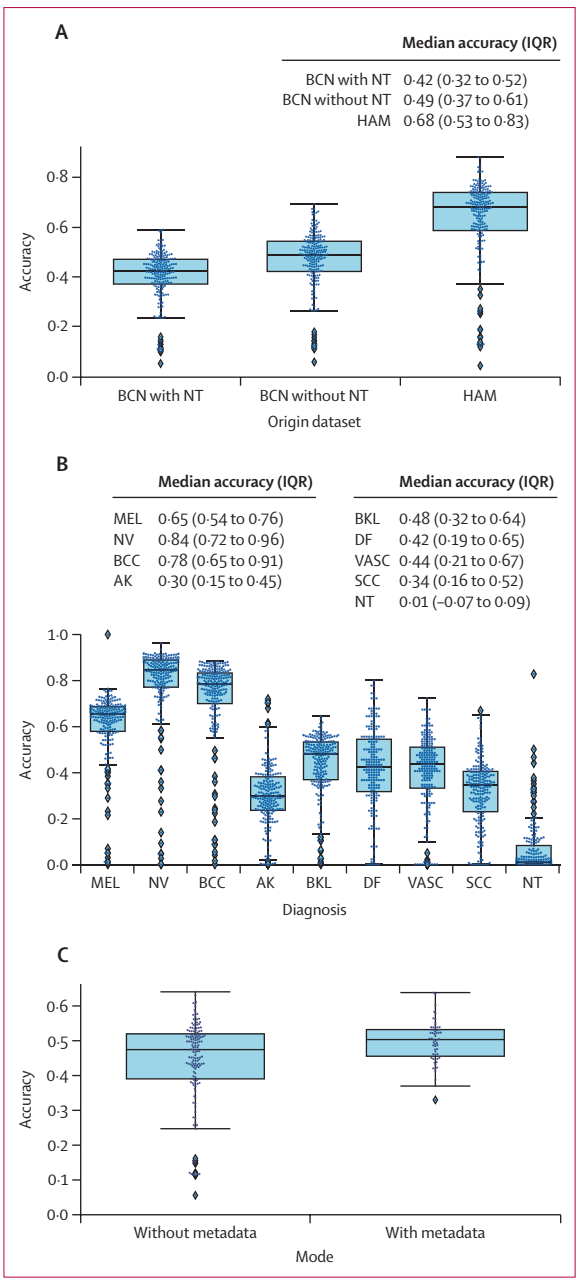


Figure 1: **Algorithm accuracy across all submissions, by dataset, metadata use, and diagnostic class**
(A) Boxplot and table showing median (IQR) for balanced accuracy across all participant submissions for each test set partition (p<0·001 for all comparisons). (B) Boxplot of diagnosis-specific balanced accuracies for each diagnostic class. (C) Comparison of balanced accuracy over all submissions with and without clinical metadata. AK=actinic keratosis. BCC=basal cell carcinoma. BCN=Hospital Clinic Barcelona. BKL=benign keratosis. DF=dermatofibroma. HAM=Medical University of Vienna. MEL=melanoma. NT=not trained. NV=nevi. SCC=squamous cell carcinoma. VASC=vascular lesions.

For **DermaChallenge** see https://dermonaut.meduniwien.ac.at/dermachallenge/

Algorithm evaluation

Challenge participants submitted a comma-separated value file to the online submission and scoring system (Covalic) containing the diagnostic predictions for each image in the test dataset. Diagnosis confidences were expressed as floating-point values in the closed interval $(0 \cdot 0, 1 \cdot 0)$.

Algorithms were ranked according to balanced multiclass accuracy (mean recall across classes after mutually exclusive classification decisions), which has the advantage of balancing for the prevalence of malignant diagnoses, especially melanoma, as compared to standard accuracy.[7] Algorithms' balanced accuracy performance was compared between data subsets using Bonferroni-adjusted paired *t* tests. The level of significance for all hypothesis tests was $0 \cdot 05$. Paired Student's *t* test was used because algorithms were evaluated on the same images. Confusion matrices and area under the receiver operating characteristic curve (AUROC), were also calculated and compared with imaging and lesion factors that each influence diagnostic accuracy (using algorithm identifiers as group labels with an exchangeable covariance matrix). Matrices are separated into nine diagnostic groups for each ground truth annotation, with aggregate statistics shown in the first row of each group (the reference row), and stratifications shown across subsequent rows. Values of the matrix convey the proportion of images with given ground truth labels (specified by group) that were assigned a particular prediction by algorithms (specified by the columns) on average across the top 25 algorithms.

Statistical analyses were performed using pandas, matplotlib, scipy, numpy, and statsmodels Python packages.[31–34]

Expert reader study

We compared the performance of the algorithms against that of dermatologists in a simulated setting that reflected intended clinical use. 18 expert board-certified dermatologists from around the world (with at least 2 years of active daily use of dermoscopy) classified images selected from a pool of 1269 images from the test set. To perform assessment, these experts (henceforth referred to as expert readers) used a custom platform, DermaChallenge, created by the Medical University of Vienna.[8,13,32,35] Expert readers were first given three training levels of 30 images each from the training dataset to practise, before classifying images from the nine diagnostic categories (including NT) in groups of 30 images at a time. To compare performance between expert readers and the algorithms, a summary sAUROC metric was used and implemented in R.[36]

Role of the funding source

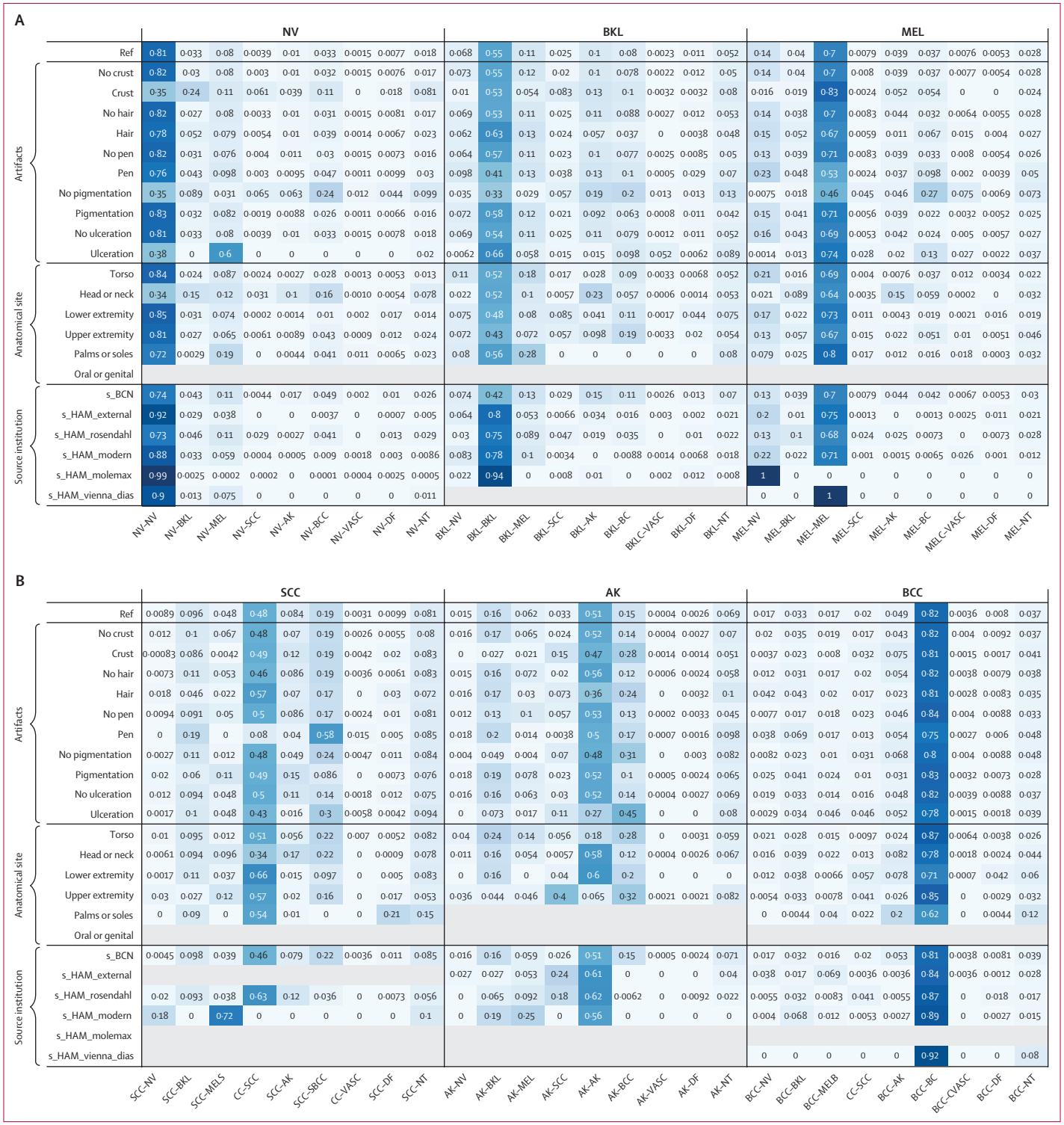
The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

169 algorithms were submitted by 64 teams, divided into the image-only task (129 submissions from 64 teams) and

the combined image and metadata task (40 submissions from 16 teams). The top two performing algorithms used ensembles of the EfficientNet architecture,[37–39] and the third-place team used ensembles of the ResNet architecture.[37] The top performing algorithm achieved 63·6% overall balanced accuracy.

The balanced algorithm accuracy on the HAM dataset partition—which is an earlier benchmark that is less



(Figure 2 continues on next page)

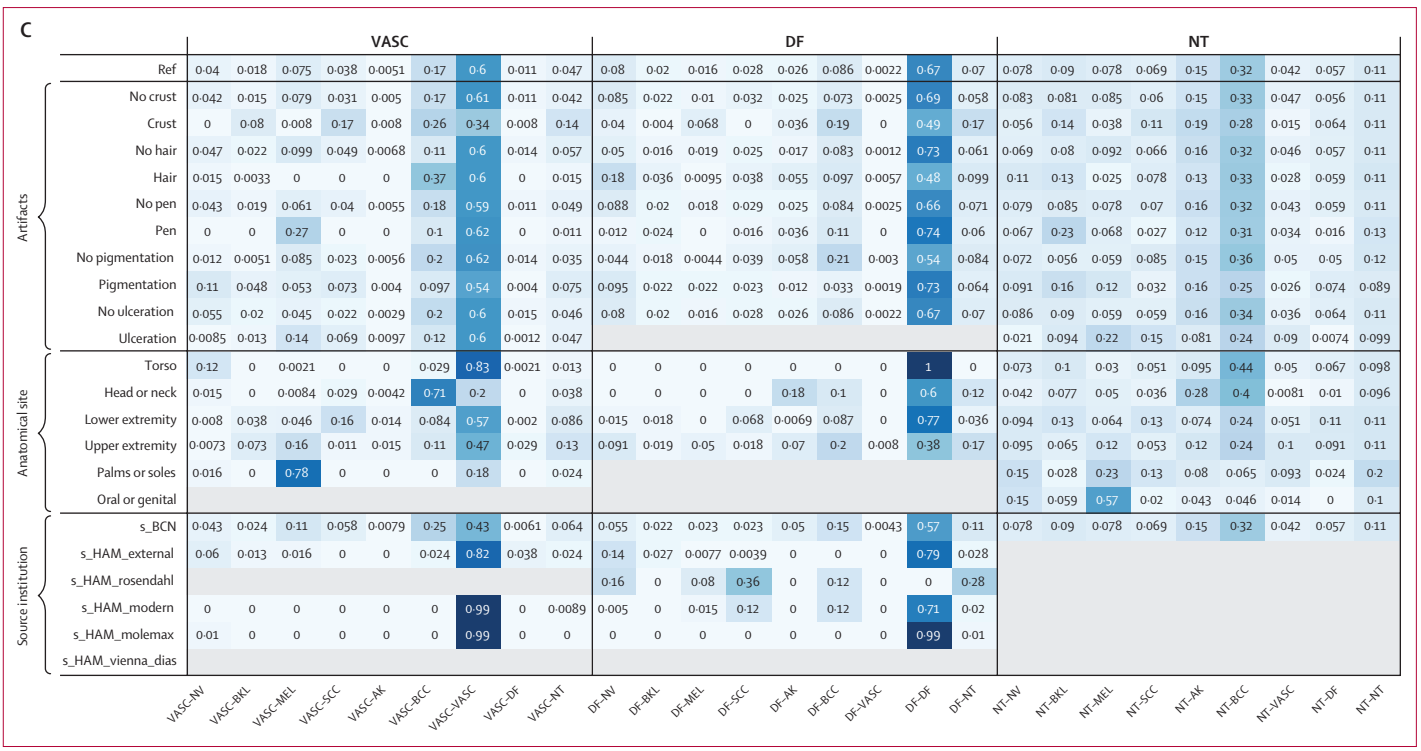


Figure 2: **Confusion matrix, separated into nine groups for each diagnostic category in the test set**

Values represent the proportion of images in the test set given a classification specified by columns, on average for the top 25 algorithms. The reference row of each group shows the aggregate values for each diagnosis. Subsequent rows include stratifications across artifacts (ie, crust, hair, pen marks), anatomical site, and source institution. Upper extremity refers to arms and feet (not palms or soles). Lower extremity refers to legs (not palms or soles). AK=actinic keratosis. BCC=basal cell carcinoma. BCN=Hospital Clinic Barcelona. BKL=benign keratosis. DF=dermatofibroma. HAM=Medical University of Vienna. MEL=melanoma. NT=not trained. NV=nevi. SCC=squamous cell carcinoma. VASC=vascular lesion.

reflective of image quality variations seen in practice—was significantly better than the BCN images, even without considering the impact of the NT category, on which all algorithms performed poorly (figure 1). Balanced accuracy of the best algorithm reduced by 23·2% (from 82·0% to 58·8%) when comparing the HAM dataset to the new images in BCN. For mean AUROC, this decrease was 0·075 (from 0·981 in HAM to 0·907 in BCN). Across all algorithms, the mean decrease in balanced accuracy between dataset partitions was 22·3% (SD 8·6; p<0·0001).

The use of auxiliary metadata (such as the lesion anatomic location, patient sex, and age) slightly improved mean algorithm accuracy from 50% (SD 15) to 56% (7; figure 1).

Across all methods, the algorithms' ability to flag the NT category was impaired relative to the algorithms' ability to classify diagnoses included in the training data (figure 2). On average across the top 25 teams, only 11% of the NT predictions were correct, which was similar to random chance (1 in 9). Most of the benign NT disease states were misclassified as basal cell carcinoma (32·4% on average across the top 25 algorithms), with another 7·8% misclassified as melanoma, and another 6·9% misclassified as squamous cell carcinoma.

The best performing team approached the NT class by training a model on external data they obtained themselves, including healthy skin, warts, cysts, and benign alterations. Other approaches used by challenge participants included direct 8-class models allowing the image not to belong to any class, and Shannon entropy estimation.[40] Despite these attempts, the top algorithm estimated only 1·6% of the NT class correctly (appendix p 6).

A confusion matrix as a function of diagnosis for the top 25 algorithms (additionally stratified according to image artifacts, anatomic site, and source institution) is shown in figure 2. The proportional representation of each category is provided in the appendix (p 3).

The influence of quantified image artifacts (such as crust, hair, or pen marks), on diagnostic accuracy is shown in subsequent rows of figure 2 across the top 25 algorithms. Diagnoses that do not frequently present with crust (such as vascular lesions, dermatofibromas, and nevi) were frequently miscategorised by the algorithms when crust was present. Presence of hair did not affect misclassification; except for actinic keratosis, where only 36% of actinic keratosis with hair present in the image were correctly classified (*vs* 56% without hair). Typically, pigmented lesions, such as nevi and melanomas, were frequently misclassified as basal cell carcinomas when they were non-pigmented (24% and 27% of the time, respectively). Typical pigmented lesions, such as nevi (83% correct when pigmented) and

melanomas (71% correct when pigmented), had decreased accuracy when non-pigmented (35% for nevi and 46% for melanomas). When non-pigmented, nevi and melanoma were frequently misclassified as basal cell carcinomas (24% and 27% of the time, respectively).

When we measured the impact of anatomical site on algorithm performance, lesions from the head and neck anatomical regions were frequently misclassified among nevi, vascular proliferations, and dermatofibromas. This finding could be a result of differences in dermoscopic patterns on skin from chronic sun damage due to their location in sun-exposed areas on the body. Regarding the impact of different image source institutions, the top 25 algorithms correctly diagnosed 99·0% of nevi correctly from s_HAM_molemax; however, no algorithms correctly identified melanoma from that same source. On average, the top 25 algorithms correctly identified 75·0% of melanomas from s_HAM_external. This disparity in diagnostic performance between image sources probably reflects the varied underlying distributions of melanomas and nevi in the datasets (appendix p 3).

The NT category was divided into five subcategories for the purpose of analysis, including scar, benign neoplasm (eg, onychomatricoma), normal variant (including hyperpigmentation and hypomelanosis), inflammatory disease (including eczema and psoriasis), and infectious disease (appendix p 3). Figure 3 presents a confusion matrix between these subcategories and other diagnostic categories included in the training data, averaged across the top 25 algorithms. Lesions that are predominantly pink, such as scars, inflammatory lesions, and benign neoplasms, were commonly misdiagnosed as basal cell carcinoma (which are also pink in colour).

We used an online interactive reader platform (DermaChallenge) to evaluate the diagnostic performance of expert readers as compared with the algorithm submissions. 82 tests of 30 images were performed (baseline distribution, table 1), each reflecting the overall distribution in the test set. This distribution was not known to the expert readers at the time of the study. We received 2460 ratings on 1269 images in rounds of 30 images each. (table 2, figure 4). The receiver operating characteristic curve analysis showed that the performance of the top three algorithms for malignancy was superior to that of expert readers, except for the NT category (figure 4). The top experts still outperformed the top three algorithms for malignancy; however, on average, experts did not outperform the top three algorithms. For the actinic keratosis diagnosis, expert readers demonstrated lower accuracy than the top three algorithms (43% *vs* 83%) but performed similarly (43% *vs* 44%) to the algorithms on average (table 2). The top three algorithms had better diagnostic accuracy than expert readers did on basal cell carcinomas (91% *vs* 70%), dermatofibromas (73% *vs* 50%), and nevi (76% *vs* 56%). Although the NT class was challenging for experts and for the algorithms, expert readers performed significantly better than all algorithms in terms of sensitivity and summary AUROC (26% correct classification *vs* 6%, p<0·0001).

	NV	BKL	MEL	SCC	AK	BCC	VASC	DF	NT
Inflammatory disease	0·055	0·072	0·025	0·095	0·19	0·36	0·02	0·081	0·10
Benign neoplasm	0·096	0·12	0·059	0·053	0·13	0·28	0·072	0·073	0·11
Normal variant	0·13	0·089	0·31	0·086	0·07	0·13	0·049	0·015	0·12
Scar	0·029	0·055	0·082	0·048	0·19	0·44	0·019	0·033	0·10
Infectious disease	0·16	0·18	0·074	0·11	0·065	0·12	0·081	0·072	0·14

Figure 3: **Confusion matrix of the diagnoses comprising the NT category**
The confusion matrix shows which of the other categories included in training the diagnoses were confused for, measured across the top 25 algorithms. AK=actinic keratosis. BCC=basal cell carcinoma. BKL=benign keratosis. DF=dermatofibroma. MEL=melanoma. NT=not trained. NV=nevi. SCC=squamous cell carcinoma. VASC=vascular lesion.

	Goal number
Actinic keratosis	1
Basal cell carcinoma	6
Benign keratosis	3
Dermatofibroma	1
Melanoma	1
Not trained	5
Nevi	8
Squamous cell carcinoma	1
Vascular lesion	1

Table 1: **Goal distribution of diagnoses included in a set of 30 images in the reader study**

Discussion

Our image classification challenge and analysis shows that, when compared with a previously published, well controlled benchmark (HAM10000), the balanced, multi-class accuracy of state-of-the-art image classification methods decreases by more than 20% on datasets specifically designed to better reflect clinical realities. Overall, a balanced accuracy of 63·6% for the top algorithm is a notable decrease in performance when compared with the previous benchmark of 86·1%.[9] We simulated intended clinical use by including images that were of varying quality, were from different source institutions, contained diagnostic categories that were not captured in the training dataset, and contained quantified imaging artifacts across both train and test datasets, all of which were found to contribute to performance degradation. Algorithms specifically designed to fail safely by flagging images outside its area of expertise were unable to complete this task. These findings highlight the poor generalisability of current state-of-the-art algorithms, and a potentially serious safety issue for clinical deployment, despite previously reported high AUROCs for malignancy on well controlled datasets.

The poor performance of algorithms on the NT category has significant implications for clinical practice. The NT

	Readers	All algorithms	Top 3 algorithms
AK*	0·43 (0·23–0·63)	0·44 (0·42–0·46)	0·83 (0·77–0·89)
BCC*	0·70 (0·61–0·79)	0·80 (0·77–0·82)	0·91 (0·88–0·95)
BKL	0·48 (0·36–0·60)	0·37 (0·35–0·39)	0·43 (0·37–0·50)
DF*	0·50 (0·30–0·71)	0·33 (0·30–0·36)	0·73 (0·50–0·95)
MEL	0·62 (0·53–0·71)	0·58 (0·56–0·60)	0·70 (0·64–0·77)
NV*	0·56 (0·46–0·66)	0·76 (0·74–0·79)	0·76 (0·74–0·77)
NT†	0·26 (0·17–0·35)	0·06 (0·05–0·08)	0·01 (0·01–0·02)
SCC	0·65 (0·46–0·83)	0·31 (0·29–0·33)	0·62 (0·55–0·69)
VASC	0·83 (0·68–0·97)	0·46 (0·43–0·49)	0·79 (0·66–0·92)

Data are accuracy of mean count (95% CI). Mean count of correct reader classifications in batches of 30 lesions was 15·7 (95% CI 14·46–16·94). Mean count of correct algorithm (best) classifications in batches of 30 lesions was 18·95 (18·20–19·70). AK=actinic keratosis. BCC=basal cell carcinoma. BKL=benign keratosis. DF=dermatofibroma. MEL=melanoma. NT=not trained. NV=nevi. SCC=squamous cell carcinoma. VASC=vascular lesion. *Top three algorithms (average) performed >20% better than readers. †Readers performed ≥20% better than algorithms.

Table 2: **Summary of reader accuracy versus that of automated classifiers**

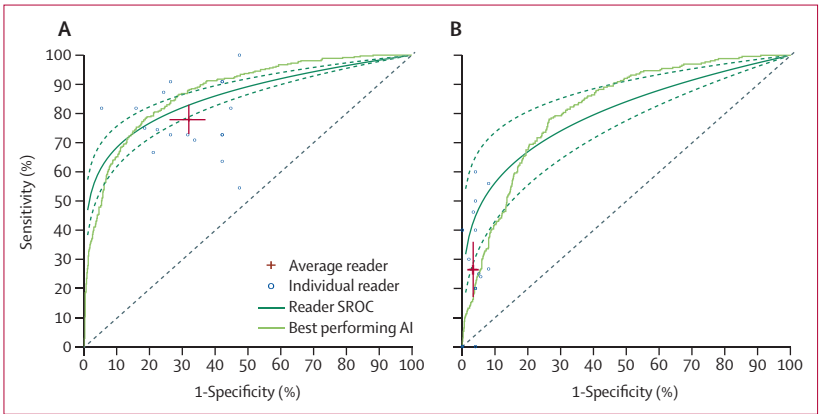


Figure 4: **Receiver operating characteristic curves for the expert readers on grouped malignant diagnoses (A) and NT class (B) as compared with the top three algorithms**

Crosses represent the average sensitivity and specificity of the readers, with the length of the bars corresponding to the 95% CI. AI=artificial intelligence. NT=not trained. SROC=summary receiver operating characteristic curve.

class was diagnosed correctly only 11% of the time across the top 25 algorithms. The NT category, which primarily comprised benign inflammatory diagnoses and scars, was confused for malignancy 47% of the time by the top 25 algorithms. NT images were most commonly confused for basal cell carcinoma, probably due to the pink colour of basal cell carcinomas and most lesions in the NT category. This leads to concerns for clinical implementation, as 47% of benign NT lesions might have been biopsied if biopsy decisions were predicated upon an automated classification system for skin lesions. In addition, false-positive malignancy predictions will contribute to patient anxiety and concern. Although the NT category was also challenging for expert clinician readers, readers performed significantly better than the algorithms (26% *vs* 6% correct, p<0·0001), on average.

Melanomas, benign keratoses, and actinic keratoses were frequently confused for one another. Clinically unusual features decreased the accuracy of algorithms' predictions compared with images without those features, such as the decrease seen between pigmented versus non-pigmented nevi (83% correct *vs* 35% correct) and melanomas (71% correct *vs* 46% correct). Source institution was also found to influence classification errors, highlighting the challenges of algorithm generalisation.

These results highlight that algorithms should be tested on both usual and unusual types of lesions and imaging attributes, and the need for algorithms with a robust capability to identify images outside of its training distributions. Caution should be used when considering the implementation of automated classification predictions into clinical workflows, especially in clinically unusual representations (such as nevi with crust, which were correctly classified in only 34·7% of cases). Careful analysis of the distribution of algorithm performance on test data according to various characteristics, such as image source, anatomical site, image attributes, and clinical features, will help stakeholders to understand how to deploy algorithms in prospective studies.

The results from our comparison of board-certified dermatologists against AI challenge submissions are consistent with previous reports. On average, the algorithms achieved higher accuracy than most expert readers (apart from the top experts who outperformed the algorithms for malignancy). However, to our knowledge this study is the first to identify a group of lesions, the NT categories, for which expert readers outperformed the automated approaches. This result exposes concerning safety issues around the deployment of automated algorithms in clinical settings, and the need to design better methods to identify images outside of an algorithm's area of expertise to avoid unnecessary biopsies or missed melanomas—both of which would have occurred if the algorithms tested in this work were deployed.

This analysis has several limitations. First, providing metadata improved algorithm predictions, but the effect size was small. This small effect size is probably due to the scarce metadata that were available for incorporation into the images. For example, it might be possible for age to be derived from the amount of sun damage visible on the background skin. Future efforts could review a more expansive list of metadata features to more deeply evaluate this impact. Second, the utility of this work is restricted by the retrospective nature of image collection, the scarce diversity in ethnicities (as presumed from clinic locations), the absence of skin tone labelling of patient images, and that the expert reader study was conducted on static images that do not mimic a clinical setting. We also included multiple lesion timepoints, which highlights the difficulty of gold standard labelling of melanomas that develop from benign neoplasms. Future work could investigate this transition to improve AI detection. Third, we tested algorithms against scenarios and statistical shifts that are highly dependent

on the training dataset. Although the specific decreases in performance we report might not be generalisable to other applications and training distributions, the considerations outlined, the image artifacts that are found to impact accuracy, and the algorithm failure on images have not been trained to recognise should be considered for all applications. There is increasing evidence that human–computer interaction might improve upon the accuracy of humans or AI alone.[15] Further work would benefit from a prospective approach to dataset design, and closely supervised trials of automated approaches with clinicians in clinical practice.

In summary, this large dermoscopic image classification challenge showed that the accuracy of state-of-the-art classification methods decreases by more than 20% on datasets specifically designed to better reflect clinical realities, as compared with a previous, well controlled benchmark. Quantified imaging artifacts contained in both training and testing datasets were found to decrease accuracy when accuracy was stratified by artifacts and disease conditions. In addition, algorithms specifically designed to fail safely by flagging images outside their training data performed worse than expert readers. These results highlight potentially serious safety issues for clinical deployment, despite previous well controlled datasets reporting high AUROCs for diagnoses such as malignancy.

Contributors

NC, AH, and JM supervised the study. AH acquired funding and handled project administration. MC, NC, and VR wrote the original draft of the manuscript, did the formal data and statistical analyses, and accessed and verified the underlying data. MC, VR, SD, HK, CR, JW, CC, BH, NRK, KL, SPo, and SPu curated data. DG, HK, MAM, CR, PT, and JW were responsible for investigation methodology. DG, HK, MAM, CR, PT, JW, CC, BH, NRK, KL, SPo, SPu, AH, and JM reviewed and edited the manuscript. All authors had access to the data presented in the manuscript. VR, NC, and MC, in collaboration with all authors, were responsible for the decision to submit for publication.

Declaration of interests

NC was an employee of IBM during a portion of this work. IBM's only involvement was through the work of NC. NC is currently employed at Microsoft, but Microsoft's only involvement is through the work of NC. NC holds two dermatology patents (US patent 10568695 B2 for surgical skin lesion removal and US patent 10255674 B2 for surface reflectance reduction in images using non-specular portion replacement) that are not relevant to this work. NC reports holding stock in IBM and Microsoft; payment or honoraria from Memorial Sloan-Kettering; and support for attending meetings or travel from Memorial Sloan-Kettering and IBM. SPu consults or receives honoraria from Almirall, ISDIN, Bristol Myers Squibb, La Roche-Posay, Pfizer, Regeneron, Sanofi, and SunPharma. VR is an expert advisor for Inhabit Brands. PT receives honoraria from Silverchair and Lilly. BH is employed by Kitware. AH consults for Canfield Scientific. All other authors declare no competing interests.

Data sharing

Individual challenge participant data are available, upon request, on the ISIC archive website. This includes all challenge submission results and manuscripts detailing all challenge submission approaches. It will be made available indefinitely to anyone who wishes to access these data. Training data for the 2019 challenge are available at https://challenge.isic-archive.com/landing/2019. Algorithms are publicly evaluable on the 2019 test data used in this study through the ISIC Challenge platform.

Acknowledgments

This work is supported by the National Institutes of Health/National Cancer Institutes Cancer Center Support Grant (P30 CA008748), the Melanoma Research Alliance Young Investigator Award (614197), La Marató de TV3 (718/C/2019), and the Charina Fund.

References

1 Cancer Today. Population fact sheets. https://gco.iarc.fr/today/fact-sheets-populations (accessed July 7, 2020).
2 Gershenwald JE, Scolyer RA, Hess KR, et al. Melanoma staging: evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin* 2017; **67:** 472–92.
3 Kim JYS, Kozlow JH, Mittal B, et al. Guidelines of care for the management of cutaneous squamous cell carcinoma. *J Am Acad Dermatol* 2018; **78:** 560–78.
4 Wolner ZJ, Yélamos O, Liopyris K, Rogers T, Marchetti MA, Marghoob AA. Enhancing skin cancer diagnosis with dermoscopy. *Dermatol Clin* 2017; **35:** 417–37.
5 Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *Br J Dermatol* 2008; **159:** 669–76.
6 Conde Taboada A, García Doval I, Feal C, et al. Distribución geográfica de dermatólogos y plazas MIR de dermatología en España. *Piel Form Contin En Dermatol* 2003; **18:** 477–80.
7 Codella N, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). *arXiV* 2018; published online Jan 8. https://arxiv.org/abs/1710.05006v3 (preprint).
8 Codella N, Rotemberg V, Tschandl P, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* 2019; published online March 29. http://arxiv.org/abs/1902.03368 (preprint).
9 Tschandl P, Codella N, Akay BN, 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.
10 Gutman D, Codella N, Celebi ME, et al. Skin lesion analysis toward melanoma detection: a challenge at the International Symposium on Biomedical Imaging 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiV* 2016; published online May 4. https://arxiv.org/abs/1605.01397 (preprint).
11 Marchetti MA, Codella NCF, Dusza SW, 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:** 270–77.e1.
12 Marchetti MA, Liopyris K, Dusza SW, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the International Skin Imaging Collaboration 2017. *J Am Acad Dermatol* 2020; **82:** 622–27.
13 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.
14 Rotemberg V, Halpern A, Dusza S, Codella NC. The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice. *Semin Cutan Med Surg* 2019; **38:** e38–42.
15 Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med* 2020; **26:** 1229–34.
16 Maron RC, Weichenthal M, Utikal JS, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *Eur J Cancer* 2019; **119:** 57–65.
17 Haenssle HA, Fink C, Toberer F, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol* 2020; **31:** 137–43.

For the **ISIC archive** see https://challenge.isic-archive.com/leaderboards/2019

18 Sun MD, Kentley J, Mehta P, Dusza S, Halpern AC, Rotemberg V. Accuracy of commercially available smartphone applications for the detection of melanoma. *Br J Dermatol* 2021; published online Nov 22. https://doi.org/10.1111/bjd.20903.

19 Liang S, Srikant R, Li Y. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv 2020; published online Aug 30. https://arxiv.org/abs/1706.02690 (preprint).

20 Mar VJ, Soyer HP. Artificial intelligence for melanoma diagnosis: how can we deliver on the promise? *Ann Oncol* 2019; **30:** e1–3.

21 Winkler JK, Fink C, Toberer F, 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.

22 Maron RC, Haggenmüller S, von Kalle C, et al. Robustness of convolutional neural networks in recognition of pigmented skin lesions. *Eur J Cancer* 2021; **145:** 81–91.

23 Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science* 2019; **363:** 1287–89.

24 Haenssle HA, Fink C, Schneiderbauer R, 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.

25 International Skin Imaging Collaboration. ISIC 2019 leaderboards. 2019. https://challenge.isic-archive.com/leaderboards/2019/ (accessed July 7, 2020).

26 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.

27 Combalia M, Codella NCF, Rotemberg V, et al. BCN20000: dermoscopic lesions in the wild. *arXiv* 2019; published online Aug 30. http://arxiv.org/abs/1908.02288 (preprint).

28 Marchetti MA, Codella NCF, Dusza SW, 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 A Acad Dermatol* 2018; **78:** 270–77.e1.

29 Marchetti MA, Liopyris K, Dusza SW, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the International Skin Imaging Collaboration 2017. *J Am Acad Dermatol* 2020; **82:** 622–27.

30 Cohn D. Active learning. In: Sammut C, Webb GI, eds. Encyclopedia of machine learning. Boston, MA: Springer US; 2010: 10–14.

31 Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. https://conference.scipy.org/proceedings/scipy2010/seabold.html (accessed Oct 1, 2020).

32 Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng* 2007; **9:** 90–95.

33 Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; **17:** 261–72.

34 Oliphant T. A guide to NumPy. Austin, TX: Continuum Press, 2006.

35 Rinner C, Kittler H, Rosendahl C, Tschandl P. Analysis of collective human intelligence for diagnosis of pigmented skin lesions harnessed by gamification via a web-based training platform: simulation reader study. *J Med Internet Res* 2020; **22:** e15597.

36 Holling H, Böhning W, Böhning D. Meta-analysis of diagnostic studies based upon SROC-curves: a mixed model approach using the Lehmann family. *Stat Model* 2012; **12:** 347–75.

37 Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, CA: AAAI Press; 2017: 4278–84.

38 Ha Q, Liu B, Liu F. Identifying melanoma images using EfficientNet ensemble: winning solution to the SIIM-ISIC Melanoma Classification Challenge. *arXiv* 2020; published online Oct 11. http://arxiv.org/abs/2010.05351 (preprint).

39 Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. *arXiv* 2020; published online Sept 11. http://arxiv.org/abs/1905.11946 (preprint).

40 Pacheco AGC, Ali A-R, Trappenberg T. Skin cancer detection based on deep learning and entropy to detect outlier samples. *arXiv* 2020; published online Jan 5. https://arxiv.org/abs/1909.04525 (preprint).