THE LANCET
Digital Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Combalia M, Codella N, Rotemberg V, et al. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. *Lancet Digit Health* 2022; **4:** e330–39.

Data and Datasets Used

Table 1 Demographic and metadata comparison between HAM and BCN subsets

	Age	Sex (male)	Anatomic Site Unknown	Anatomic Site torso	Anatomic Site lower extremity	Anatomic Site upper extremity	Anatomic Site head/neck	Anatomic Site palms/soles	Anatomic Site oral/genital
HAM	51.36 +- 17.61	7558 (51.68 %)	2989 (20 %)	5177 (35 %)	3152 (21 %)	1765 (12 %)	1522 (10 %)	18 (0 %)	0 (0 %)
BCN	57.39 +- 18.22	9840 (51.93 %)	297 (1 %)	7369 (38 %)	3616 (19 %)	1934 (10 %)	4961 (26 %)	670 (3 %)	99 (0 %)

We used a combination of two datasets to train and validate submitted models. Training datasets included the HAM10000 dataset (1,2) and a curated dataset from Hospital clinic Barcelona (3).

The dataset from Hospital Clinic Barcelona (BCN) was developed by retrospective search of the clinical database for images that could be attached to diagnosis either by histopathology (for malignancy and excised benign cases) or by expert review and clinical follow up. Due to the nature of the study, it is a convenience sample that was intended to be as large as possible due to well described need for large datasets to improve AI model development.

Comparison of demographic data and metadata comparison between HAM and BCN subsets is shown in Table 1 for all images (including for development and validation). Image artifacts were only classified in the test dataset, and are shown in Table 2. As shown in Table 2, there were proportionately more image artifacts such as crust, pen, and ulceration in the BCN subset. Hair was similarly represented in both subsets. Pigmentation was more commonly represented in the HAM dataset, but this was likely to decrease performance in the BCN subset for melanoma detection and increase confusion around nonpigmented lesions. Proportional representation across diagnoses of the image artifacts that were found to affect AI performance in the test dataset is shown in Table 3.

Table 2 Comparison of image artifacts present in HAM vs BCN subsets (of the test datasets only)

dataset	crust	hair	pen	pigmentation	ulceration
HAM	19 (1%)	350 (23%)	0 (0%)	1440 (95%)	29 (1%)
BCN	562 (9%)	1147 (19%)	1104 (18%)	3966 (68%)	496 (8%)

Unknown categories are described in further detail in Table 6. Spanish pathology clinic notes and pathology reports were reviewed and translated by native speakers of Spanish. Two expert dermatologists reviewed the categorization and agreed on the categorizations. Algorithms were only scored on the overall NT category not on the subcategories. Excluded categories were malignancies, or lesions that were found to be in the 8 trained categories such as seborrheic keratoses, but are maintained here as they are still available in the test dataset as used for the challenge.

Distribution of diagnostic classes between HAM and BCN subsets are shown in Table 3.

Table 3 *Distribution of diagnostic classes between HAM and BCN subsets (combined train and test)*

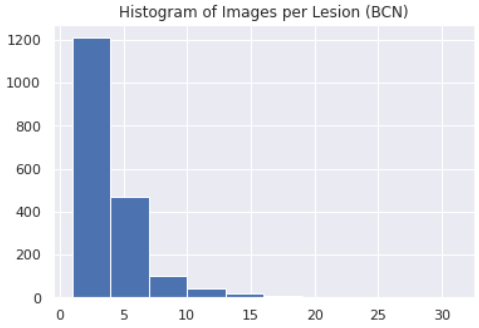
	NV	MEL	BCC	UNK	BKL	AK	SCC	VASC	DF
HAM	9701 (66 %)	1857 (12 %)	622 (4 %)	0 (0%)	1725 (11 %)	149 (1 %)	229 (1 %)	180 (1 %)	160 (1 %)
BCN	5669 (29 %)	3992 (21 %)	3676 (19 %)	2047 (10 %)	1559 (8 %)	1092 (5 %)	564 (2 %)	177 (0 %)	170 (0 %)

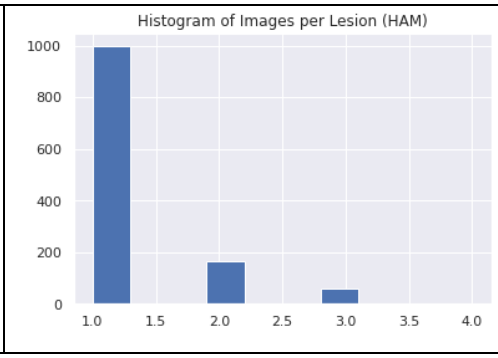
All biopsied lesions had their gold standard labeled histopathologically. Borderline lesions on histopathology were excluded from this analysis due to challenges with gold standard labeling of even histopathologically evaluated cases. However, further work needs to be done to improve gold standard labeling of intermediate cases in the future and this would be a rich area for investigation. Percent of overall lesions confirmed by histopathology is shown in Table 4. Lesion IDs corresponding to images will be available upon request.

Multiple images were allowed per lesion. Comparison of multiple images per lesion is shown in Figure 1.

Table 4 *Proportion of histopathologically confirmed cases*

	Histopatholgy
HAM	786 (51 %)
BCN	4646 (71 %)





AK	4.16
BCC	2.43
BKL	2.19
DF	1.64
MEL	4.75
NV	1.97
SCC	2.62
NT	3.06
VASC	2.06

Figure 1 Histogram of images per lesion in the subsets BCN (left), HAM (middle) and a table of the average number of images per lesion by diagnostic class

	AK	BCC	BKL	DF	MEL	NV	SCC	NT	VASC
crust	7 % (28)	17 % (161)	8 % (50)	11 % (10)	1 % (17)	1 % (34)	31 % (48)	16 % (228)	5 % (5)
hair	24 % (88)	17 % (159)	17 % (105)	23 % (21)	14 % (170)	25 % (591)	16 % (25)	22 % (309)	24 % (24)
pen	45 % (170)	31 % (293)	12 % (79)	11 % (10)	7 % (82)	17 % (401)	5 % (8)	4 % (54)	7 % (7)
pigmentation	79 % (294)	53 % (507)	88 % (560)	70 % (63)	94 % (1159)	97 % (2299)	35 % (55)	31 % (439)	30 % (30)
ulceration	3 % (12)	11 % (109)	2 % (13)	0 % (0)	12 % (145)	0 % (2)	31 % (48)	12 % (163)	33 % (33)
anterior torso	10 % (37)	43 % (416)	25 % (156)	2 % (2)	29 % (357)	42 % (1003)	43 % (67)	21 % (290)	15 % (15)
head/neck	81 % (302)	28 % (267)	32 % (205)	2 % (2)	21 % (253)	7 % (169)	29 % (46)	33 % (468)	19 % (19)
lower extremity	1 % (2)	13 % (121)	15 % (93)	32 % (29)	19 % (231)	20 % (481)	15 % (24)	25 % (355)	20 % (20)
oral/genital	0 % (0)	0 % (0)	0 % (0)	0 % (0)	0 % (0)	0 % (0)	0 % (0)	2 % (28)	0 % (0)
palms/soles	0 % (0)	1 % (9)	0 % (1)	0 % (0)	9 % (117)	2 % (55)	3 % (4)	5 % (76)	5 % (5)
posterior torso	1 % (2)	2 % (15)	6 % (37)	1 % (1)	4 % (47)	7 % (170)	1 % (2)	0 % (0)	4 % (4)
upper extremity	5 % (19)	9 % (82)	8 % (49)	28 % (25)	15 % (188)	8 % (185)	8 % (12)	11 % (148)	11 % (11)
BCN	95 % (355)	90 % (865)	66 % (418)	51 % (46)	86 % (1063)	62 % (1463)	85 % (133)	100 % (1399)	65 % (66)
HAM_external	1 % (3)	3 % (33)	12 % (79)	34 % (31)	3 % (32)	5 % (120)	0 % (0)	0 % (0)	18 % (18)
HAM_Rosendahl	3 % (13)	3 % (29)	9 % (59)	1 % (1)	4 % (55)	4 % (88)	14 % (22)	0 % (0)	0 % (0)
HAM_ViDIR_Current	1 % (3)	3 % (30)	9 % (59)	9 % (8)	6 % (80)	9 % (223)	1 % (2)	0 % (0)	9 % (9)
HAM_ViDIR_MoleMax	0 % (0)	0 % (0)	3 % (20)	4 % (4)	0 % (1)	19 % (462)	0 % (0)	0 % (0)	8 % (8)
HAM_ViDIR_Legacy	0 % (0)	0 % (1)	0 % (0)	0 % (0)	0 % (3)	1 % (15)	0 % (0)	0 % (0)	0 % (0)

Percentages may not add to 100: values close to zero were rounded to 0% and unlabeled values are not listed (such as for anatomic site)

Table 5 Proportion of images that contain features found to affect diagnostic misclassification (all images with those features not just those that were misclassified)

Table 6 Categorization of unknown categories, as translated from clinic notes and pathology reports in their original Spanish

Benign neoplasm	Benign Neoplasm Cyst Onychomatricoma Lymphoid infiltrate Vascular proliferation Hyperkeratosis Dermatofibroma Melanocytic proliferation Hidrocystoma Sebaceous adenoma Fibroma Acantholytic keratoma Cutaneous Horn
Exclude	Nonprimary melanoma Merkel cell carcinoma Other Nevus Seborrheic Keratosis Lymphoma Lentigo Paget's disease
Scar	Scar
Infectious disease	Infectious disease Impetigo Mycosis
Normal variant	Skin pigmentation Normal skin Ulcer Nail Acanthosis Hyperplasia Melanonychia Hair Erosion Hypomelanosis Hematoma
Inflammation	Inflammatory Dermatitis Granuloma Inflammation Capillaritis Panniculitis Lichen simplex Drug reaction Pseudodegos Necrobiosis Mastocytosis Rosacea Insect bite Prurigo Eczema Morphea Sweet Syndrome

Reader Study

We tasked 22 expert readers with analyzing groups of 30 images at a time for multiclass labels. We compared the best reader, the average reader, the winning algorithm (according to balanced accuracy), and the average algorithm. While we consider balanced accuracy to be the best metric for comparison since it was the main outcome measure and the criteria upon which the algorithms were judged, we also compared sensitivity, specificity, and Area under the receiver operating curve (AUC). Demographics for the readers are shown in Table 7. Readers were compared only to the without-metadata task, Task 1 and were not given metadata due to space constraints on the smartphone screen.

Table 7 Reader study demographics

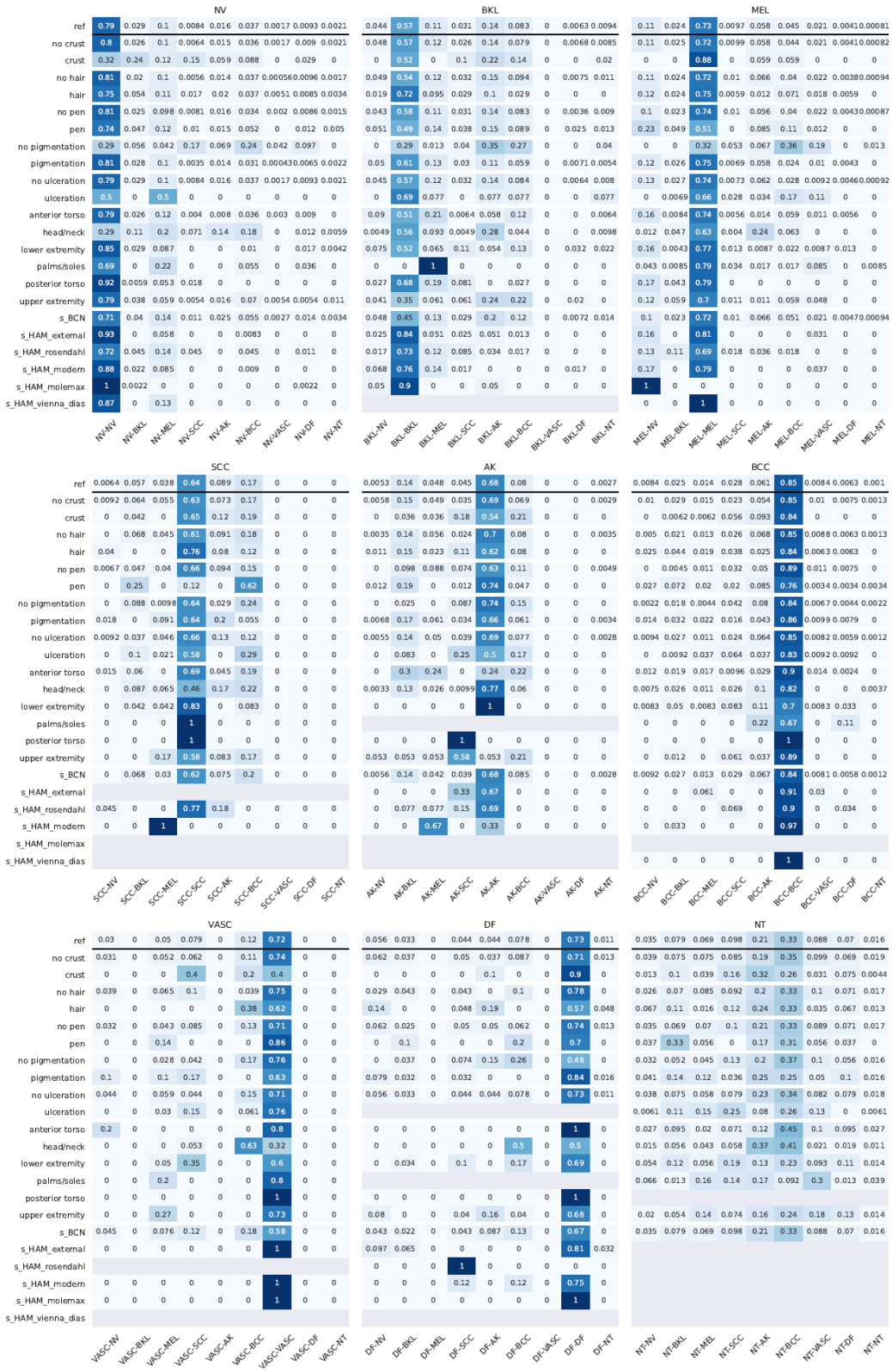
Category		Count
Age		
	22-31	2
	32-41	13
	42-51	3
	52-61	3
	62+	1
Years Experience		
	0-1	5
	3	2
	5	9
	10	6
Gender		
	Male	13
	Female	9

AUC were calculated for readers using summary ROCs described (4). On average, the readers outperformed the average algorithm in balanced accuracy, sensitivity for malignancy, and sensitivity for out of distribution images. The best reader outperformed the best algorithm across all metrics. The best algorithm outperformed the average reader across all metrics for malignancy and balanced accuracy, but not for classification of not trained (NT) images.

	Balanced Multiclass Accuracy	Classifying Malignancy			Classifying Not Trained Images		
		Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
Average Reader	58.0% (54%-63%)	77.9% (73%-83%)	68.1% (63%-74%)	0.86 (0.83-0.89)	26.5% (17%-36%)	96.7% (96%-98%)	0.8 (0.73-0.88)
Best Reader	80.8%	81.2%	84.2%	N/A	40.0%	100.0%	N/A
Average algorithm (n=83) - (merged) multiclass labels	43.7% (41%-46%)	70.0% (66%-74%)	70.2% (68%-72%)	0.79 (0.76-0.80)	5.1% (3%-8%)	97.3% (95%-99%)	0.6 (0.61-0.66)
Winning algorithm (DaisyLab 1418) - (merged) multiclass labels	63.8%	86.0%	69.8%	0.88	1.1%	99.9%	0.81

The "Best Reader" was defined as the reader who achieved the highest overall accuracy in the reader study.
95% CI for average reader BMCA, sensitivity, and specificity are the 2.5th and 97.5th percentiles from boostrapped replicates of 18 readers' BMCA, sensitivity, and specificity metrics.
95% CI for average reader AUC from the uncertainty in the estimation of "theta" using Holling et. al. approach for estimating sROC for meta-analysis of diagnostic studies (4).
95% CI for average algorithm BMCA, sensitivity, and specificity are the 2.5th and 97.5th percentiles from boostrapped replicates of 83 algorithms' BMCA, sensitivity, and specificity metrics.
95% CI for average algorithm AUC are the 2.5th and 97.5th percentiles from boostrapped replicates of 48 algorithms' AUC metrics, because 48 of 83 provided multiclass outputs on a continous scale.

Performance of the top team submission across all categories is shown below:



References

1. 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.

2. 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(7):938–47.

3. Combalia M, Codella NCF, Rotemberg V, Helba B, Vilaplana V, Reiter O, et al. BCN20000: Dermoscopic Lesions in the Wild. ArXiv190802288 Cs Eess [Internet]. 2019 Aug 30 [cited 2020 Jul 7]; Available from: http://arxiv.org/abs/1908.02288

4. 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 Aug 1;12(4):347–75.