Evaluation of Artificial Intelligence–Assisted Diagnosis of Skin Neoplasms: A Single-Center, Paralleled, Unmasked, Randomized Controlled Trial

Seung Seog Han[1,2,6], Young Jae Kim[3,6], Ik Jun Moon[3,6], Joon Min Jung[3], Mi Young Lee[3], Woo Jin Lee[3], Chong Hyun Won[3], Mi Woo Lee[3], Seong Hwan Kim[4], Cristian Navarrete-Dechent[5] and Sung Eun Chang[3]

Trial design: This was a single-center, unmasked, paralleled, randomized controlled trial. **Methods:** A randomized trial was conducted in a tertiary care institute in South Korea to validate whether artificial intelligence (AI) could augment the accuracy of nonexpert physicians in the real-world settings, which included diverse out-of-distribution conditions. Consecutive patients aged >19 years, having one or more skin lesions suspicious for skin cancer detected by either the patient or physician, were randomly allocated to four nondermatology trainees and four dermatology residents. The attending dermatologists examined the randomly allocated patients with (AI-assisted group) or without (unaided group) the real-time assistance of AI algorithm (https://b2020.modelderm.com#world; convolutional neural networks; unmasked design) after simple randomization of the patients. **Results:** Using 576 consecutive cases (Fitzpatrick skin phototypes III or IV) with suspicious lesions out of the initial 603 recruitments, the accuracy of the AI-assisted group (n = 295, 53.9%) was found to be significantly higher than those of the unaided group (n = 281, 43.8%; $P = 0.019$). Whereas the augmentation was more significant from 54.7% (n = 150) to 30.7% (n = 138; $P < 0.0001$) in the nondermatology trainees who had the least experience in dermatology, it was not significant in the dermatology residents. The algorithm could help trainees in the AI-assisted group include more differential diagnoses than the unaided group (2.09 vs. 1.95 diagnoses; $P = 0.0005$). However, a 12.2% drop in Top-1 accuracy of the trainees was observed in cases in which all Top-3 predictions given by the algorithm were incorrect. **Conclusions:** The multiclass AI algorithm augmented the diagnostic accuracy of nonexpert physicians in dermatology.

Journal of Investigative Dermatology (2022) **142,** 2353–2362; doi:10.1016/j.jid.2022.02.003

INTRODUCTION

With advancement of deep learning algorithms, promising results have been reported in the diagnosis of skin cancer, but most of the studies were retrospective (Esteva et al., 2017; Haenssle et al., 2020; Han et al., 2020c; Maron et al., 2021; Tanaka et al., 2021; Tschandl et al., 2019). A relatively small number of prospective studies have been reported in machine learning research (Liu et al., 2019; Topol, 2020). Among them, only 11 randomized controlled trials (RCTs) were published in 2021 (Zhou et al., 2021), and to the best of our knowledge, no RCT studies have been published in dermatology. Unlike retrospective studies, the cases of a prospective study include untrained diseases (out-of-distribution), and the results are affected by the quality of photographs and the expertise of the user.

In the field of dermatology, only a small number of prospective (non-RCT) studies have been reported (Dascalu and David, 2019; Muñoz-López et al., 2021; Navarrete-Dechent et al., 2021). A commercial dermoscopy algorithm showed 88.1% sensitivity and 78.8% specificity, compared with teledermoscopists (MacLellan et al., 2021). An algorithm using dermoscopic lens attachments showed the ability to identify melanoma with an accuracy similar to that of specialists (Phillips et al., 2019). Finally, the performance of onychomycosis algorithm that outperformed all 42 dermatologists on the receiver operating characteristic curve in a retrospective study (Han et al., 2018) was found to be equivalent to that of 5 dermatologists in the prospective study (Kim et al., 2020).

We previously developed a unified multiclass skin disease classifier (Model Dermatology; https://modelderm.com) and showed that the algorithm could classify 134 skin disorders at

[1]*Department of Dermatology, I Dermatology Clinic, Seoul, Korea;* [2]*IDerma, Inc, Seoul, Korea;* [3]*Department of Dermatology, Asan Medical Center, Ulsan University College of Medicine, Seoul, Korea;* [4]*Department of Plastic and Reconstructive Surgery, Kangnam Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Korea; and* [5]*Department of Dermatology, School of Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile*

[6]*These authors contributed equally to this work.*

Correspondence: Sung Eun Chang, Department of Dermatology, Asan Medical Center, Ulsan University College of Medicine, 88, OLYMPIC-RO 43-GIL Songpa-gu, Seoul 05505, Korea. . E-mail: csesnumd@gmail.com

Abbreviations: AI, artificial intelligence; GP, general practitioner; RCT, randomized controlled trial

Received 3 November 2021; revised 26 January 2022; accepted 8 February 2022; accepted manuscript published online 18 February 2022; corrected proof published online 2 June 2022

Table 1. Demographics and the Status of the Randomization

Variable	Unaided Group (n = 281)	AI Group (n = 295)	Overall (n = 576)
Age (y)	58.6 ± 18.0	58.7 ± 18.0	58.6 ± 18.0
Sex (male)	48.4% (136)	41.4% (122)	44.8% (258)
Fitzpatrick skin phototype			
Type III	75.1% (211)	79.0% (233)	77.1% (444)
Type IV	24.9% (70)	21.0% (62)	22.9% (132)
Race	All Asian	All Asian	All Asian
Onset (y)[1]	4.3 ± 7.7	5.7 ± 8.6	5.0 ± 8.2
Size (mm)	10.9 ± 11.0	9.8 ± 9.6	10.3 ± 10.3
Recent changes			
Size	54.1% (152)	48.8% (144)	51.4% (296)
Color	14.2% (40)	14.6% (43)	14.4% (83)
Shape	11.7% (33)	14.6% (43)	13.2% (76)
Site			
Head and neck	41.6% (117)	43.7% (129)	42.7% (246)
Trunk	23.1% (65)	22.0% (65)	22.6% (130)
Arm	15.3% (43)	12.2% (36)	13.7% (79)
Leg	19.9% (56)	22.0% (65)	21.0% (121)
Family history of skin cancer	2.1% (6)	1.7% (5)	1.9% (11)
Suspected by patients	45.2% (127/277)	47.1% (139/295)	46.2% (266/576)
Pathologically diagnosed cases	91.8% (258)	90.2% (266)	91.0% (524)
Malignancy	16.0% (45)	13.2% (39)	14.4% (83)
Benign	75.8% (213)	76.9% (227)	76.6% (441)
Clinically-diagnosed cases	8.2% (23)	9.8% (29)	9.0% (52)
Trainees			
Non-DER trainees	49.1% (138)	50.8% (150)	50.0% (288)
Dermatology residents	50.9% (143)	49.2% (145)	50.0% (288)
Participated period (day)	18.4 ± 14.3	17.6 ± 13.8	18.0 ± 14.1
Participated No. of cases	40.9 ± 26.6	39.1 ± 25.9	40.0 ± 26.2
Attending dermatologists			
Experience after the board certification (y)	12.4 ± 8.9	11.4 ± 8.6	11.9 ± 8.8
Use of dermoscopy	19.2% (54)	20.0% (59)	19.6% (113)

Abbreviations: AI, artifical intelligence; DER, dermatology; No., number.
[1]A total of 88.6% (249 cases), 90.2% (266 cases), and 89.4% (515 cases) onset records were available in the unaided group, AI group, and the overall, respectively.

a dermatology resident level (Han et al., 2020c). However, in a prospective study using 340 consecutive teledermatology cases (Muñoz-López et al., 2021), Top-1 accuracy of the algorithm (41.2%) was lower than that of the general practitioners (49.3%), probably because 10.3% of the teledermatology cases belonged to the untrained classes (out-of-distribution).

In this study, we performed a single-center, paralleled, unmasked, RCT to investigate whether a multiclass artificial intelligence (AI) algorithm can instantly improve the accuracy (primary outcome) and sensitivity/specificity (secondary outcome) of nondermatologists who examined patients with suspicious skin neoplasms detected by either the patient or the physician.

RESULTS

This study is reported per the Consolidated Standards of Reporting Trials—Artificial Intelligence guidelines (Liu et al., 2020). Top-(n) accuracy is the accuracy of the Top-(n) diagnoses. If any one of the Top-(n) diagnoses is correct, it counts as 'correct'.

AI group versus unaided group

To evaluate that the two groups were truly comparable, accuracies of the attending dermatologists and trainees (before interventions) were compared. Top-1 accuracy of attending dermatologists (62.3%) and trainees (43.8%) of the unaided group were higher than those of the AI group (dermatologists = 60.3%, trainees = 41.0%), which indicated that easier cases were not disproportionately allocated to the AI group (Table 1 and Figure 1).

Overall, the Top-1 accuracy of the AI group was 53.9% and that of the unaided was 43.8% (all trainees; chi-square test, $P = 0.019$; Figure 2 and Table 2). There were significant differences in the result depending on whether the participant was a nondermatology trainee (general practitioner [GP]) or a dermatology resident (resident). The Top-1 accuracy of the AI_{GP} group (54.7%) was markedly higher than that of unaided$_{GP}$ group (29.7%; chi-square test, $P < 0.0001$), whereas the Top-1 accuracy of the $AI_{resident}$ group and the unaided$_{resident}$ group was 53.1% and 57.3%, respectively (chi-square test, $P = 0.55$). In the AI group, we compared the judgment before and after receiving assistance of the algorithm, and there was a significant enhancement in the Top-1

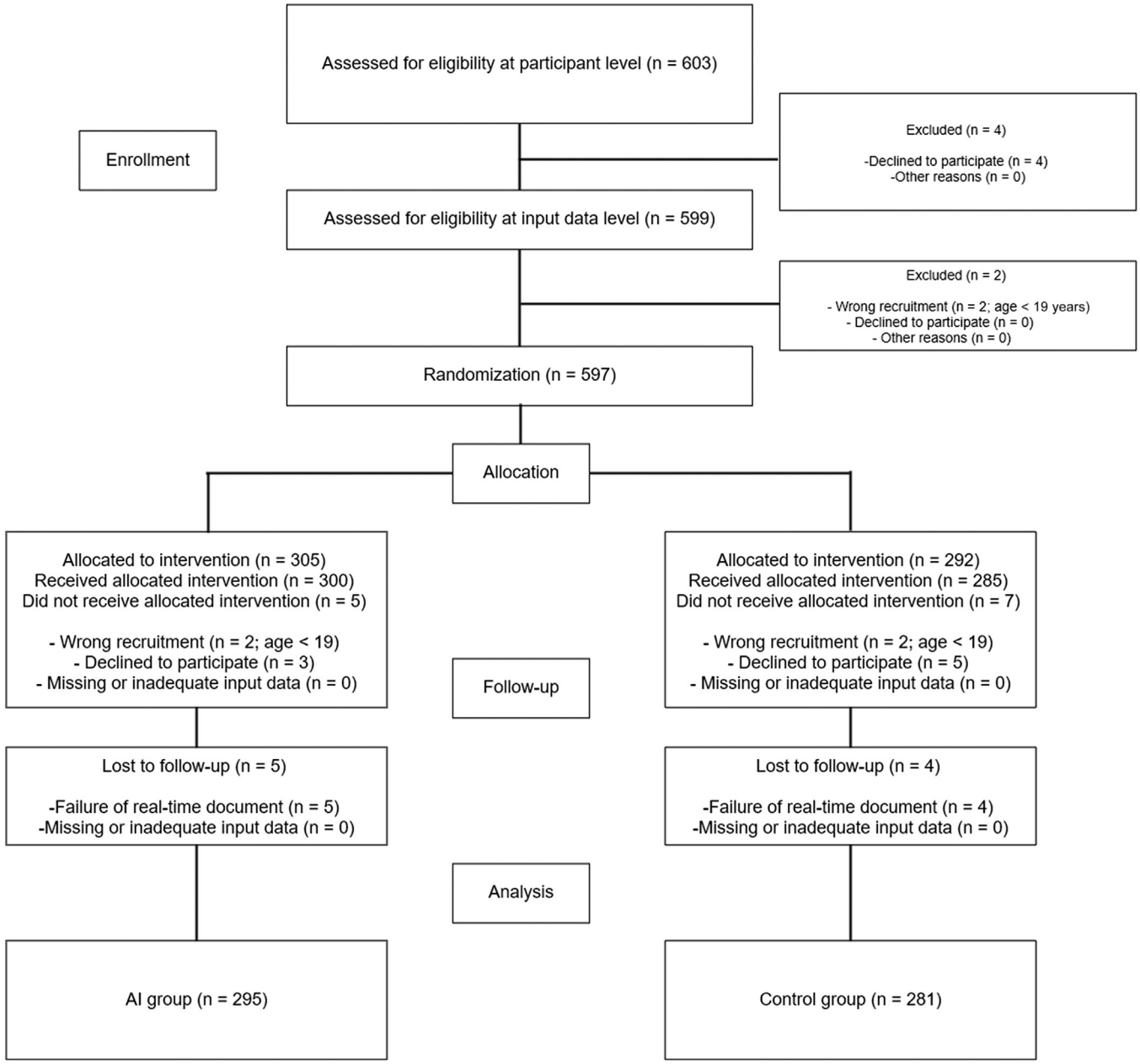


Figure 1. Flowchart. AI, artificial intelligence.

accuracy of the AI_{GP} group (before augmentation = 30.7%, after augmentation = 54.7%; McNemar test, $P < 0.0001$). However, in the $AI_{resident}$ group, the change was not significant (before augmentation = 51.7%, after augmentation = 53.1%; McNemar test, $P = 0.86$). As shown in Supplementary Table S1, a greater improvement in accuracy was observed for the subsequent cases, up to +25.0% and +37.5% for Top-1 and Top-3 accuracy, respectively, although the accuracy for Top-1 and Top-3 was improved by only +0.0% and +15.0%, respectively, for the first five cases.

When the analysis was restricted to 266 cases that were biopsied, the Top-1/Top-3 accuracy of the standalone algorithm, trainees before augmentation, trainees after augmentation, and attending dermatologists were 51.5%/75.9%, 40.2%/49.2%, 56.0%/71.4%, and 56.0%/66.2%, respectively. The accuracies of the AI-augmented trainees were equivalent to those of the attending dermatologists. In the 258 biopsy-proven cases of the unaided group, the Top-1/Top-3 accuracy of trainees and attending dermatologists were 42.6%/57.4% and 58.9%/68.6%, respectively.

Malignancy determination affects clinical decisions such as ordering a biopsy if there is any malignancy among Top-3 predictions. Based on the Top-3 predictions, the sensitivity/specificity of the AI group and the unaided were 84.6%/69.5% and 75.6%/62.7%, respectively (chi-square test, $P = 0.45/0.13$; Table 2). The sensitivity/specificity of the AI_{GP} group and the $unaided_{GP}$ group were 80.0%/81.5% and 56.3%/68.9%, respectively (chi-square test, $P = 0.24/0.029$). The sensitivity/specificity of the $AI_{resident}$ group and the $unaided_{resident}$ group were 89.5%/57.1% and 86.2%/56.1%, respectively (chi-square test, $P = 1.0/0.98$).

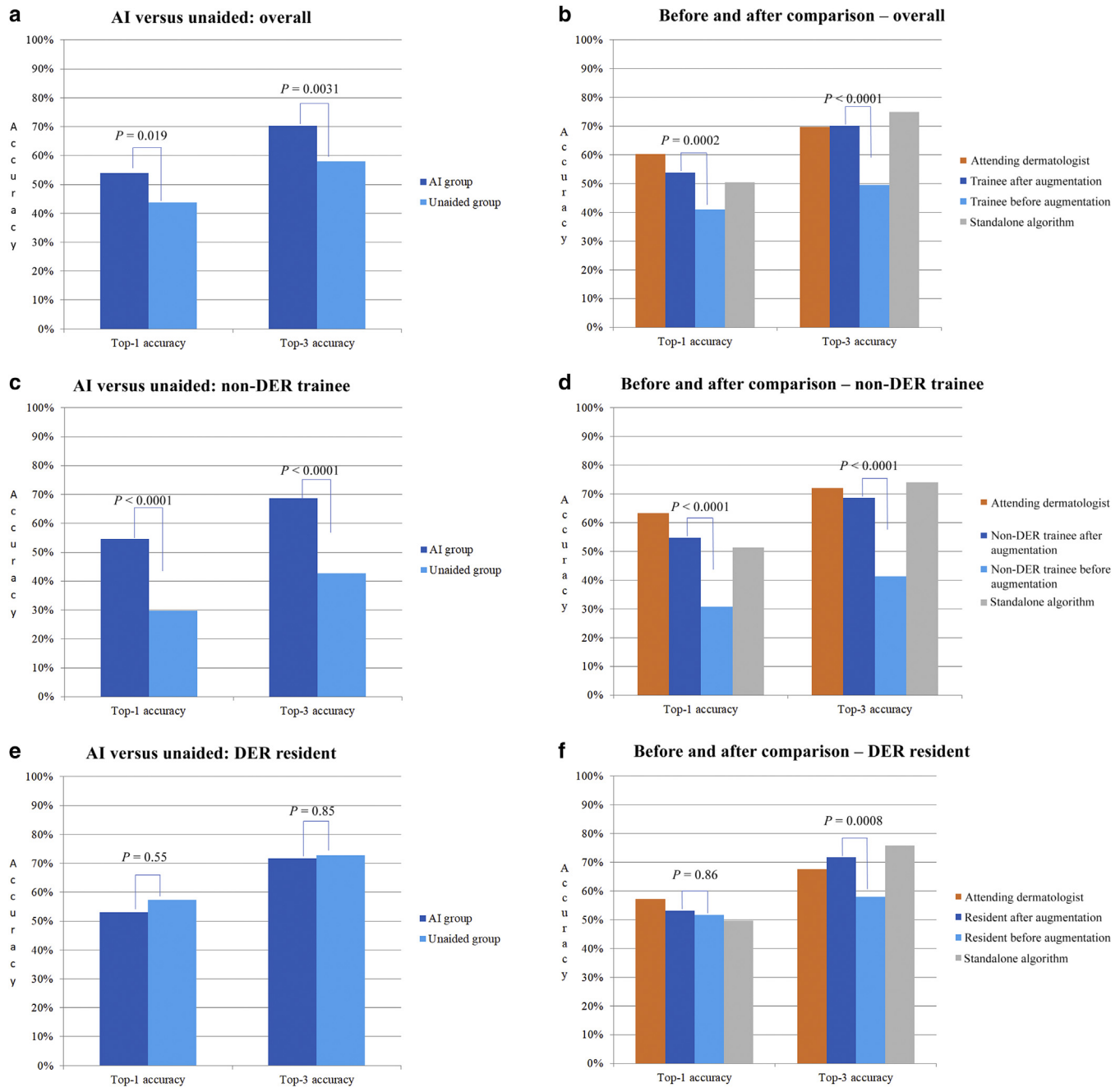


Figure 2. Top accuracies for diagnosing exact diseases. (**a**) All trainees: AI (n = 295) versus unaided (n = 281). (**b**) All trainees: before and after the assistance of the algorithm in the AI group (n = 295). (**c**) Non-DER trainees: AI (n = 150) versus unaided (n = 138). (**d**) Non-DER trainees: before and after the assistance of the algorithm in the AI group (n = 150). (**e**) DER residents: AI (n = 145) versus unaided (n = 143). (**f**) DER residents: before and after the assistance of the algorithm in the AI group (n = 145). Overall, eight trainees participated (four dermatology residents and four nondermatology trainees). The Top accuracy stratified by 82 disorders is described in Supplementary Table S6. The *P*-values of Top accuracies between the AI group and the unaided group and between before and after the assistance of the trainees are described in Table 2. AI, artificial intelligence; DER, dermatology.

There was a significant difference in the number of differential diagnoses between the AI group and the unaided group (2.09 vs. 1.95; Wilcoxon rank-sum test, $P = 0.0005$). The number of differential diagnoses of the AI_{GP} group (2.00) was higher than that of the $unaided_{GP}$ group (1.88; Wilcoxon rank-sum test, $P < 0.0001$). The number of differential diagnoses of the $AI_{resident}$ group (2.17) was also higher than that of the $unaided_{resident}$ group (2.01; Wilcoxon rank-sum test, $P = 0.019$).

Before and after comparison: Individual analysis

The Top-1 and Top-3 accuracy before the use of AI was 41.3% ± 15.7% and 49.2% ± 13.2%, respectively (8 trainees with 295 cases; Supplementary Table S2). The Top-1 and Top-3 accuracy after the assistance of AI was increased to 53.6% ± 9.3% and 69.8% ± 10.6%, respectively.

Individual improvement in Top-1 diagnostic accuracy for each trainee ranged from 5.4% to +41.4%, with an average (SD) of +12.2% (16.6%), which was not statistically

Table 2. Accuracy, Sensitivity, and Specificity of the Participants and Algorithm

	AI Group				Unaided Group		Before Versus After Augmentation		AI Versus Unaided Group	
Performance Index	Trainee Before Augmentation	Trainee After Augmentation	Standalone AI	Attending Dermatologist	Trainee	Attending Dermatologist	Difference (95% CI)[1]	*P*-Value[2]	Difference (95% CI)[1]	*P*-Value[3]
	All trainees (295 cases, 8 trainees)				All Trainees (281 cases, 8 trainees)					
Accuracy (Top-1)	41.0% (121/295)	53.9% (159/295)	50.5% (149/295)	60.3% (178/295)	43.8% (123/281)	62.3% (175/281)	+12.9% (+4.9%−+20.9%)	0.0002	+10.1% (+2.0%−+18.3%)	0.019
Accuracy (Top-3)	49.5% (146/295)	70.2% (207/295)	74.9% (221/295)	69.8% (206/295)	58.0% (163/281)	71.2% (200/281)	+20.7% (+12.9%−28.4%)	<0.0001	+12.2% (+4.4%−+19.9%)	0.0031
Sensitivity from Top-1	64.1% (25/39)	64.1% (25/39)	66.7% (26/39)	71.8% (28/39)	53.3% (24/45)	64.4% (29/45)	0.0% (−21.3% to +21.3%)	1.00	+10.8% (−10.2% to +31.7%)	0.44
Specificity from Top-1	85.9% (220/256)	90.6% (232/256)	90.6% (232/256)	91.8% (235/256)	85.2% (201/236)	91.5% (216/236)	+4.7% (−1.0% to +8.7%)	0.059	+5.5% (−0.3% to +11.2%)	0.085
Sensitivity from Top-3	82.1% (32/39)	84.6% (33/39)	84.6% (33/39)	89.7% (35/39)	75.6% (34/45)	88.9% (40/45)	+2.6% (−14.0% to +19.1%)	1.00	+9.1% (−7.8% to +26.0%)	0.45
Specificity from Top-3	60.9% (156/256)	69.5% (178/256)	61.3% (157/256)	67.2% (172/256)	62.7% (148/236)	66.5% (157/236)	+8.6% (+0.4% − +16.8%)	0.017	+6.8% (−1.5% to 15.2%)	0.13
	Non-DER Trainees (150 Cases, 4 Trainees)				**Non-DER Trainees (138 Cases, 4 Trainees)**					
Accuracy (Top-1)	30.7% (46/150)	54.7% (82/150)	51.3% (77/150)	63.3% (95/150)	29.7% (41/138)	63.8% (88/138)	+24.0% (+13.1% to +34.9%)	<0.0001	+25.0% (+13.9% to +36.0%)	<0.0001
Accuracy (Top-3)	41.3% (62/150)	68.7% (103/150)	74.0% (111/150)	72.0% (108/150)	42.8% (59/138)	71.0% (98/138)	+27.3% (+16.5% to +38.2%)	<0.0001	+25.9% (+14.8% to +37.0%)	<0.0001
Sensitivity from Top-1	65.0% (13/20)	55.0% (11/20)	65.0% (13/20)	85.0% (17/20)	43.8% (7/16)	50.0% (8/16)	+10.0% (−20.2% to +40.2%)	0.77	+11.3% (−21.4% to +43.9%)	0.74
Specificity from Top-1	85.4% (111/130)	94.6% (123/130)	92.3% (120/130)	91.5% (119/130)	83.6% (102/122)	95.1% (116/122)	+9.2% (+2.0%− +16.4%)	0.014	+11.0% (+3.4%− +18.6%)	0.0088
Sensitivity from Top-3	75.0% (15/20)	80.0% (16/20)	80.0% (16/20)	90.0% (18/20)	56.3% (9/16)	93.8% (15/16)	+5.0% (−20.8% to +30.8%)	1.00	+23.8% (−6.2% to +53.7%)	0.24
Specificity from Top-3	70.0% (91/130)	81.5% (106/130)	65.4% (85/130)	72.3% (94/130)	68.9% (84/122)	67.2% (82/122)	+11.5% (+1.2%− +21.9%)	0.025	+12.7% (+2.1%− +23.3%)	0.029
	DER Residents (145 Cases, 4 Trainees)				**DER Residents (143 Cases, 4 Trainees)**					
Accuracy (Top-1)	51.7% (75/145)	53.1% (77/145)	49.7% (72/145)	57.2% (83/145)	57.3% (82/143)	60.8% (87/143)	+1.4% (−10.1% to +12.9%)	0.86	+4.2% (−7.2% to 15.7%)	0.55
Accuracy (Top-3)	57.9% (84/145)	71.7% (104/145)	75.9% (110/145)	67.6% (98/145)	72.7% (104/143)	71.3% (102/143)	+13.8% (+2.9% − +24.7%)	0.0008	+1.0% (−9.3% to +11.3%)	0.95
Sensitivity from Top-1	63.2% (12/19)	73.7% (14/19)	68.4% (13/19)	57.9% (11/19)	58.6% (17/29)	72.4% (21/29)	+10.5% (−18.8% to +39.9%)	0.68	+15.1% (−11.6% to +41.8%)	0.45
Specificity from Top-1	86.5% (109/126)	86.5% (109/126)	88.9% (112/126)	92.1% (116/126)	86.8% (99/114)	87.7% (100/114)	0.0% (−8.4% to +8.4%)	1.00	−0.3% (−8.9% to +8.3%)	1.00
Sensitivity from Top-3	89.5% (17/19)	89.5% (17/19)	89.5% (17/19)	89.5% (17/19)	86.2% (25/29)	86.2% (25/29)	0.0% (−19.5% to +19.5%)	1.00	+3.3% (−15.4% to +21.9%)	1.00
Specificity from Top-3	51.6% (65/126)	57.1% (72/126)	57.1% (72/126)	61.9% (78/126)	56.1% (64/114)	65.8% (75/114)	+5.6% (−6.7% to +17.8%)	0.34	+1.0% (−11.6% to +13.6%)	0.98

Abbreviations: AI, artificial intelligence; CI, confidence interval; DER, dermatology.

[1]The 95% CIs for two independent binomial proportions were calculated using the Wald method (wald2ci of the PropCIs package; R version 4.1.1).

[2]McNemar test was performed for the paired values.

[3]Chi-square test was performed. Positive predictive value and negative predictive value are listed in the Supplementary Table S8.

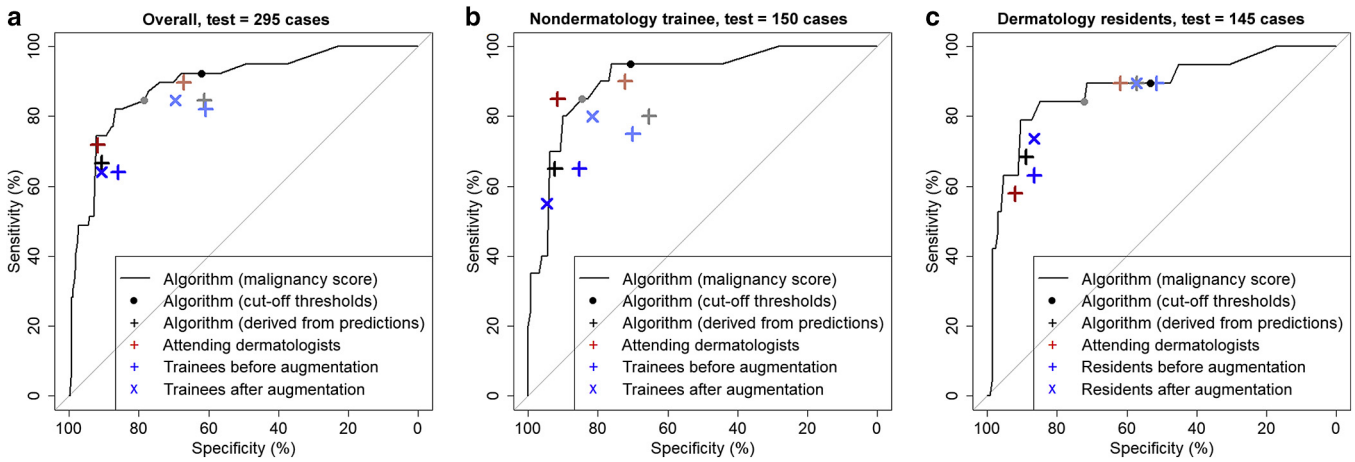


Figure 3. Sensitivity and specificity on the ROC curve for determining malignancy in the AI group. (**a**) All trainees in the AI group (n = 295). (**b**) Nondermatology trainees in the AI group (n = 150). (**c**) Dermatology residents in the AI group (n = 145). Dark blue cross (+): trainees before augmentation; malignancy decision derived from Top-3 predictions. Pale blue cross (+): trainees before augmentation; malignancy decision derived from Top-1 prediction. Dark blue x-cross (×): trainees after augmentation; malignancy decision derived from Top-3 predictions. Pale blue x-cross (×): trainees after augmentation; malignancy decision derived from Top-1 prediction. Dark red cross (+): attending dermatologists; malignancy decision derived from Top-3 predictions. Pale red cross (+): attending dermatologists; malignancy decision derived from Top-1 prediction. Black cross (+): algorithm; malignancy decision derived from Top-3 predictions. Pale black cross (+): algorithm; malignancy decision derived from Top-1 prediction. Black line: algorithm; malignancy decision derived from the malignancy score. Black dot (●): algorithm at the high-sensitivity threshold. Pale black dot (●): algorithm at the high-specificity threshold. AI, artificial intelligence; ROC, receiver operating characteristic.

significant (paired *t*-test after Shapiro-Wilk normality test, $P = 0.076$). On the contrary, Top-3 accuracy was improved by +0%−+41.4%, with an average improvement of +20.6% (12.9%), which was statistically significant (paired *t*-test after Shapiro-Wilk normality test, $P = 0.0027$). There was individual variation in the degree of improvement: the Top-1 accuracy of one dermatology resident decreased (−5.3%), whereas that of one nondermatology trainee was improved by +41.4%.

Standalone performance of the algorithm

The standalone Top-1 and Top-3 accuracy of the algorithm in the AI group was 50.5% and 74.9%, respectively. Area under the curve for determining malignancy was 0.894 (95% confidence interval = 0.838−0.950; DeLong method), which was equivalent to that of the attending dermatologists on the receiver operating characteristic curve (Figure 3). At the high-sensitive threshold, the sensitivity and specificity were 92.3% and 62.1%, respectively, and at the high-specificity threshold, the sensitivity and specificity were 84.6% and 78.5%, respectively. The sensitivity/specificity of the standalone algorithm derived from the Top-1 was 66.7%/90.6% and that from the Top-3 was 84.6%/61.3% (Table 2).

Adverse effects of the incorrect predictions from the algorithm

A 12.2% drop in Top-1 accuracy of the trainees was observed in cases where all Top-3 predictions from the algorithm were incorrect (Supplementary Tables S3 and S4). In out-of-distribution cases where Top-3 predictions from the algorithm were inevitably incorrect, the Top-1 accuracy of the trainees dropped by 5.6%.

Both sensitivity/specificity derived from Top-3 predictions of trainees were negatively affected (before augmentation = 84.6%/60.7% and after augmentation = 61.5%/59.0%) when all Top-3 predictions from the algorithm were incorrect.

There were four cases of malignancy (three basal cell carcinomas and one Bowen's disease) for which the trainees had initially included a malignant condition in their Top-3 but ruled out malignancy in their final Top-3 recorded after the use of the algorithm. Of these four cases, two were predicted to be conditions requiring follow-up visits (dysplastic nevus and actinic keratosis).

DISCUSSION

In this RCT analyzing 576 cases, we showed that a multiclass AI algorithm helped improve the diagnostic accuracy of the trainees. The augmentation was significant in nondermatology trainees who had only minimal experience in dermatology, whereas the augmentation was nonsignificant in dermatology residents. Regarding the standalone performance with 266 biopsied cases, the accuracies of the AI-augmented trainees were comparable to those of the attending dermatologists. In addition, the standalone algorithm using the malignancy score showed comparable performance to attending dermatologists in determining malignancy. This is a unique result because this study was conducted in the real-world settings, which included diverse out-of-distribution conditions.

Although several retrospective studies have showed successful results regarding the diagnosis of skin lesions using AI algorithms, most were carried out in experimental settings (Haggenmüller et al., 2021). Various factors, including but not limited to those listed in this study, make these promising results unlikely to be reproduced in real-life.

1. Clever-Hans type bias (Lapuschkin et al., 2019): the predictions of algorithms may be drawn from hidden features with no relevance, especially if the amount of training data is small. However, it is very difficult for researchers to

check whether the Clever-Hans bias exists during a retrospective experiment.
2. The presence of out-of-distribution in training classes: algorithms have no diagnostic ability at all on untrained diseases. Although our old algorithm showed a dermatologist-level of performance with the in-distribution 134 disorders (Han et al., 2020c), the performance deteriorated in the prospective study (Muñoz-López et al., 2021) with consecutive patients having diverse disorders, which indicates the relevance of the out-of-distribution problem. Although algorithms are trained on rare diseases, the diagnostic ability may be poor because only a small amount of training data for these rare diseases is available for the training.
3. The presence of out-of-distribution in characteristics: in retrospective experiments, cases with typical features are selected, whereas cases with atypical morphology are usually dropped out. Moreover, ideal photographs in terms of quality and composition are usually included in the test, which does not well represent the cases in the real world. In a prospective study with consecutive cases, an algorithm may show uncertainty to all kinds of out-of-distribution cases.
4. Fit disease prevalence of training dataset: accuracy can be optimized according to the disease prevalence of the training dataset. A model may be prone to predict disorders with high prevalence to achieve high accuracy in the internal validation, rather than learning the disease features.
5. Unpaired comparison (Genin and Grote, 2021): dermatologists do not make clinical diagnosis based solely on photographs. The clinicians in the real world use all clinical inputs (i.e., clinically history, touch, body distribution), and having no access to other diagnostic input is thus an unpaired comparison. In most circumstances, history taking and physical examination significantly improve the physician's diagnostic ability except for the mass diagnostic tests such as mammography.

Although algorithms outperformed dermatologists in previous retrospective studies, they may perform inferior to dermatologists in real-world prospective studies. In this study, suspected skin neoplasms were selected as an intended use because the performance of the algorithm for skin neoplasms was better than that of dermatologists in the previous reader tests (Han et al., 2020a, 2020c), and most kinds of neoplastic disorders were in-distribution. In addition, acquiring proper composition does not require dermatological knowledge and could be standardized. Nevertheless, in this study, the standalone Top-1 accuracy of the algorithm (50.5%) was inferior to that of the attending dermatologists (60.3%) in the real-world settings.

Dermatology does not merely deal with visual assessment of skin lesions. Although algorithms can outperform dermatologists in reader tests, it does not mean that the algorithms outperform dermatologists in real-world settings. In a cohort study with 43 skin tumors, the accuracy of the algorithm was superior to that of the dermatologists in the reader test (49.5% vs. 37.7%) but inferior to the attending physicians who examined the patients in person (68.1% vs. 49.5%) (Han et al., 2020a). The importance of in-person examination was also shown in a study using dermoscopic images in which the diagnostic accuracy of the reader test was lower than that of the physicians who actually performed the dermoscopic evaluation (Dinnes et al., 2018).

In this study, there was a marked improvement in the accuracy of the nondermatology trainees having only minimal experience in dermatology. This finding is similar to what has been reported in a previous article (Tschandl et al., 2020). Another interesting finding is that the augmented accuracy of the trainees (Top-1/Top-3 = 56.0%/71.4%) was equivalent to that of attending physicians (Top-1/Top-3 = 56.0%/66.2%) for the 266 biopsied cases. The accuracies of both the standalone algorithm and trainees were lower than that of the attending dermatologists, but synergy was found in the AI-augmented trainees. All potential diagnoses presented by the algorithm were reviewed by the trainees capable of performing a physical examination and history taking, which may result in the synergy.

With the current technology, improving the accuracy and reducing biases of algorithms require a huge amount of data. It may be better for humans to understand the diagnostic strengths and limitations of the AI and to adapt to the diagnostic characteristics of the machines. The performance of the algorithm using the malignancy score was equivalent to that of the attending dermatologists for determining malignancy on the receiver operating characteristic curve (Figure 3). At the high-sensitivity cut-off threshold, the malignancy score showed 92.3% sensitivity that can compensate for the low sensitivity (66.7%) derived from the Top-1 prediction. However, trainees did not show synergy in the binary determination as much as they did in the augmented accuracy (multiclass classification). In addition, there was no increase in Top-1 accuracy in the first five cases, whereas improved Top-1 accuracy was observed in the following cases, meaning that it took time for the participants to make full use of the algorithm. If the participants had a better understanding of the characteristics of the algorithm, the results could have been further improved. Therefore, detailed instructions on the diagnostic characteristics of algorithms should be provided for the users to improve diagnostic performance in future studies (Daneshjou et al., 2021).

There are certain limitations of this study that need to be addressed. First, this study was designed to represent the actual clinical settings in a primary care clinic. We intended to show how a general physician, with scarce experience in dermatology, could benefit from an algorithm when deciding whether to refer a patient to a tertiary hospital. However, the experimental conditions in which this study was carried out were not totally identical to the actual settings in primary care clinics. Because this study was performed prospectively in a tertiary care center, the referred cases could not represent all diverse benign conditions seen in primary care clinics. In addition, the intended user (primary care physicians) may not even think about using the algorithm owing to the lack of dermatology knowledge. In addition, if the intended user is proficient in using other tools such as dermoscopy, the algorithm may not be of great assistance for those experienced physicians.

Second, although there was a time gap between the training of the algorithm and the test, there may be a hidden bias (e.g. Clever-Hans type [Lapuschkin et al., 2019]) that we were not aware of because the clinical images, previously collected from Asan Medical Center, were used for the training.

Third, validation was conducted only in Asians, mostly with Fitzpatrick skin phototypes III and IV. To enable generalizability, further prospective studies should be performed because disease prevalence, subtype distribution, and visual characteristics of disorders may differ between ethnicities, countries, and regions. The retrospective results of the algorithm using the Edinburgh dataset from the white population (1,300 images; Top-1/Top-3 accuracy = 65.2%/84.8%, area under the curve for determining malignancy = 0.937; Supplementary Figure S1 and Supplementary Table S5) need to be validated in prospective studies.

Fourth, owing to the limited number of physicians who took part in this study (n = 8), a slight variation in individual performance may cause a biased overall result. To minimize the effect of individual variation and enhance representativeness of the results, future studies with a larger number of participants are warranted.

Finally, in determining malignancy, there was no statistical difference in the sensitivity and specificity between the AI group and the unaided group except for the specificity of the nondermatology trainees. This suggests the need for further clinical evaluation of the algorithm in terms of malignancy detection. In addition, only nine melanoma cases were included in this study. Because melanoma prevalence is relatively low in skin phototypes III and IV, future studies including other skin phototypes and more melanoma cases seem necessary.

In conclusion, our algorithm could enhance the accuracy of nonspecialists in diagnosing suspected cutaneous neoplasms in real-world settings. The assistance of the algorithm was the greatest for the least experienced physicians. In addition, the algorithm could help nondermatology trainees include more diagnoses in the list of differentials. To further improve the algorithm's performance, combining additional input data (e.g., metadata, dermoscopic images) with lesion images seems necessary. Further larger multicenter studies in various regions and ethnicities are required to validate the generalizability of our results.

MATERIALS AND METHODS

This prospective study was approved by the Institutional Review Board of Asan Medical Center (Seoul, Korea) (S2018-1703-0001). It was performed in the Department of Dermatology at Asan Medical Center, a tertiary care center in Korea. The study was conducted from 30 November 2020 to 9 September 2021 after online registration (KCT0005614; cris.nih.go.kr). The development of the algorithm (Model Dermatology, Build2020; https://b2020.modelderm.com#world) is described in the Supplementary Materials, and the algorithm was fixed on 19 September 2020. Alongwith the prediction of five differential diagnoses, the algorithm reports a malignancy score (range = 0–100). The malignancy score was defined as the sum of malignant outputs and 0.2 × premalignant outputs as used previously (Han et al., 2020b). Using the subset of the Seoul National University dataset (240 images; https://doi.org/10.6084/m9.figshare.6454973), the high-sensitivity threshold for determining malignancy was defined as the threshold at which 90% sensitivity was obtained because the sensitivity of the attending dermatologists was at the level of 88.1% (Han et al., 2020a). The high-specificity threshold was defined as the threshold at which 80% sensitivity was obtained.

In a previous pilot study (Kim et al., 2022), the Top-1 accuracy of trainees was 47.9%. If 25% enhancement after the assistance were regarded as significant, the sample size was calculated as 548 (alpha = 0.05 and power = 0.8), and we planned to recruit 600 cases (Rosner, 2011).

All patients signed an informed consent before inclusion in the study. We included consecutive patients (aged >19 years) who had one or more skin lesions suspicious for skin cancer, detected by either the patient or physician. Exclusion criteria of patients and input data included patient refusal (10 cases), wrong recruitment (6 cases; aged ≤19 years), biopsy refusal (2 cases), and nonreal-time analysis (9 cases). Broken blindness and disclosure of the biopsy results in the referral note were also in the exclusion criteria, but there was no such case, and there was no performance error for the loss of internet connection or other technical issues. Formal pathologic diagnosis (504 cases) was used as the ground truth; however, if the pathologic report consisted of a pathologic description only (i.e., lichenoid reaction), the pathologic diagnosis was determined by clinicopathological correlation (20 cases). Clinical diagnosis of the attending dermatologists was used as the ground truth for the 52 cases where biopsy was not performed because the attending dermatologists decided not to biopsy the definitely benign cases. Ultimately, 524 biopsy-proven cases and 52 clinically-diagnosed cases were included in the final analysis among the 603 cases of the initial recruitment (Table 1). As shown in Supplementary Tables S6 and S7, a total of 53 conditions were within the trained 178 classes (in-distribution) and 29 conditions were not trained by the algorithm (out-of-distribution). There were no cases in which train-test contamination could be concerned, although there were 15 cases with photographs taken previously.

A total of four attending physicians (3, 4, 6, and 22 years of experience after board certification), four first-year dermatology residents, and four nondermatology trainees (doctors in their first year after getting a medical license who rotate between various departments) participated in this study. Attending physicians routinely recorded their impressions after thorough examinations. After the simple randomization using a custom randomizer (rand function of the PHP language) by the attending dermatologists, the trainee took the patient's medical history, performed physical examinations, took photographs, and recorded their diagnostic hypothesis in real time.

In the AI group, trainees captured and selected 1–3 photographs with age and sex metadata as an input data and uploaded them to http://b2020.modelderm.com#world using internet browsers. Then the after-diagnoses was recorded, referring to the five diagnoses suggested by the algorithm and the malignancy score (Supplementary Figure S2). The photographs that the trainees thought to be of adequate quality and composition by themselves were uploaded. Trainees were instructed so that only macroimages having the large lesion of interest at the center are uploaded. In the unaided group, trainees performed routine examinations and recorded the three most probable diagnoses, without the assistance of the algorithm. The use of dermoscopy was not allowed for all trainees.

Statistical analysis

In calculating the Top accuracy, only the exact diagnosis was recorded as correct, but giving a specific subtype of the disease was also counted to be correct. For example, intradermal nevus was counted correct for the ground truth of junctional nevus. We lumped together 364 diagnoses in natural language into the 82 diagnosis codes (https://doi.org/10.6084/m9.figshare.16640257). For evaluating a malignancy prediction, the physicians diagnoses were transformed into either malignant or benign. Top accuracies, sensitivities, and specificities were compared using Pearson's chi-square test with Yates' continuity correction (AI group vs. unaided group) or McNemar test (before vs. after the assistance of the algorithm) using R version 4.1.1, and $P < 0.05$ was statistically significant.

Data availability statement

The raw data of this study is available at https://doi.org/10.6084/m9.figshare.16640257. The diagnosis in natural language, converted diagnosis code (ALIAS sheet), the cases determined by clinicopathological correlation, accuracy/sensitivity/specificity/positive predictive value/negative predictive value of the participants and algorithm, and demographics of subjects are listed. The algorithm is accessible at https://b2020.modelderm.com/#world or https://modelderm.com (Build2020) for academic purposes, and only the images with the permission of the original owner should be uploaded.

ORCIDs

Seung Seog Han: http://orcid.org/0000-0002-0500-3628
Young Jae Kim: http://orcid.org/0000-0001-9841-5797
Ik Jun Moon: http://orcid.org/0000-0002-1123-4166
Joon Min Jung: http://orcid.org/0000-0003-3432-8306
Mi Young Lee: http://orcid.org/0000-0002-3991-4390
Woo Jin Lee: http://orcid.org/0000-0002-0549-464X
Chong Hyun Won: http://orcid.org/0000-0003-1997-2240
Mi Woo Lee: http://orcid.org/0000-0003-4669-9454
Seong Hwan Kim: http://orcid.org/0000-0001-6831-5621
Cristian Navarrete-Dechent: http://orcid.org/0000-0003-4040-3640
Sung Eun Chang: http://orcid.org/0000-0003-4225-0414

CONFLICT OF INTEREST

During the period of the study, SSH founded IDerma, Inc, and he is the chief executive officer and chief technical officer of the company. The remaining authors state no conflict of interest.

ACKNOWLEDGMENTS

We would like to thank the patients who participated in the trial. SSH and YJK had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.

AUTHOR CONTRIBUTIONS

Conceptualization: SSH, CND, SEC; Data Curation: SSH, YJK, IJM, MYL; Formal Analysis: SSH; Investigation: SSH, YJK, IJM, JMJ, WJL, CHW, MWL, SEC; Methodology: SSH, SHK, CND, SEC; Project Administration: CND, SEC; Resources: YJK, IJM, JMJ, MYL, WJL, CHW, MWL, SHK, SEC; Software: SSH; Supervision: CND, SEC; Visualization: SSH; Writing - Original Draft Preparation: SSH, YJK, IJM; Writing - Review and Editing: IJM, CND, SEC

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at www.jidonline.org, and at https://doi.org/10.1016/j.jid.2022.02.003.

REFERENCES

Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. JAMA Dermatol 2021;157:1362–9.

Dascalu A, David EO. Skin cancer detection by deep learning and sound analysis algorithms: a prospective clinical study of an elementary dermoscope. EBioMedicine 2019;43:107–13.

Dinnes J, Deeks JJ, Chuchu N, Ferrante di Ruffano L, Matin RN, Thomson DR, et al. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. Cochrane Database Syst Rev 2018;12:CD011902.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks [published correction apperas in Nature 2017;546:686] Nature 2017;542:115–8.

Genin K, Grote T. Randomized controlled trials in medical AI: a methodological critique. Philosophy of Medicine 2021;2.

Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, 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.

Haggenmüller S, Maron RC, Hekler A, Utikal JS, Barata C, Barnhill RL, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. Eur J Cancer 2021;156:202–16.

Han SS, Moon IJ, Kim SH, Na JI, Kim MS, Park GH, et al. Assessment of deep neural networks for the diagnosis of benign and malignant skin neoplasms in comparison with dermatologists: a retrospective validation study. PLoS Med 2020a;17:e1003381.

Han SS, Moon IJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. JAMA Dermatol 2020b;156:29–37.

Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PLoS One 2018;13:e0191493.

Han SS, Park I, Eun 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 2020c;140:1753–61.

Kim YJ, Han SS, Yang HJ, Chang SE. Prospective, comparative evaluation of a deep neural network and dermoscopy in the diagnosis of onychomycosis [published correction appears in PloS One 2020;15:e0244899] PLoS One 2020;15:e0234334.

Kim YJ, Na JI, Han SS, Won CH, Lee MW, Shin JW, et al. Augmenting the accuracy of trainee doctors in diagnosing skin lesions suspected of skin neoplasms in a real-world setting: a prospective controlled before-and-after study. PLoS One 2022;17:e0260895.

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

Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis [published correction appears in Lancet Digit Health 2019;1:e334] Lancet Digit Health 2019;1:e271–97.

Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ 2020;370:m3164.

MacLellan AN, Price EL, Publicover-Brouwer P, Matheson K, Ly TY, Pasternak S, et al. The use of noninvasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. J Am Acad Dermatol 2021;85:353–9.

Maron RC, Schlager JG, Haggenmüller S, von Kalle C, Utikal JS, Meier F, et al. A benchmark for neural network robustness in skin cancer classification. Eur J Cancer 2021;155:191–9.

Muñoz-López C, Ramírez-Cornejo C, Marchetti MA, Han SS, Del Barrio-Díaz P, Jaque A, et al. Performance of a deep neural network in teledermatology: a single-centre prospective diagnostic study. J Eur Acad Dermatol Venereol 2021;35:546–53.

Navarrete-Dechent C, Liopyris K, Marchetti MA. Multiclass artificial intelligence in dermatology: progress but still room for improvement. J Invest Dermatol 2021;141:1325–8.

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

Rosner B. Fundamentals of biostatistics. 7th ed. Boston, MA: Brooks/Cole; 2011.

Tanaka M, Saito A, Shido K, Fujisawa Y, Yamasaki K, Fujimoto M, et al. Classification of large-scale image database of various skin diseases using deep learning. Int J Comput Assist Rad Surg 2021;16:1875–87.

Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med 2020;26:1318–20.

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.

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.

Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. npj Digit Med 2021;4:154.

SUPPLEMENTARY MATERIALS

The training history of our algorithm (Model Dermatology; https://modelderm.com) was described previously (Han et al., 2020a, 2020b, 2018a, 2018b; Muñoz-López et al., 2021; Navarrete-Dechent et al., 2021, 2018). First, the algorithm was trained using 12 benign and malignant nodules for classification of the most common skin neoplasm (Han et al., 2018a). Because several benign disorders can mimic skin neoplasms, the algorithm should be a unified classifier that can predict a large number of disorders (Han et al., 2020b). The ASAN and Web datasets were mainly used for training the convolutional neural networks. The ASAN dataset was assembled with 120,780 clinical images acquired from 2003 to 2016 at the Department of Dermatology at Asan Medical Center (Seoul, Korea). The Web dataset consisted of images obtained using a Python script (https://github.com/whria78/skinimagecrawler), and 100−500 images per disease were downloaded using two search engines (google.com and bing.com) and manually annotated based on the image findings. Furthermore, because numerous trivial conditions may result in uncertainty, a large training dataset for the algorithm was created with the assistance of region-based convolutional neural networks (Han et al., 2020a). The algorithm was trained not only with typical lesions but also with various lesions generated with the assistance of a region-based convolutional neural network to reduce false positives. A total of 4,204,323 images crops were used and only horizontal flip was applied for the augmentation. Using PyTorch (https://pytorch.org; version 1.6), we trained our convolutional neural network models using a transfer learning method with ImageNet pretrained models. Histogram normalization was performed as a preprocessing step before training the models. The output values of SENet (Hu et al., 2018), SE-ResNeXt-101, SE-ResNeXt-50, ResNeSt-101 (Zhang et al., 2020[1]), and ResNeSt-50 were arithmetically averaged to obtain a final model output. The hyperparameters were set as follows: learning_rate = 0.001, gamma = 0.1, weight_decay = 0.00001, mini_batch_size = 32, solver = SGD, momentum = 0.9, total_iteration = 90 epoch, and step_iteration = 30 epoch. As a validation set, the subset of the ASAN dataset (17,125 images of nodular disorders) was used, and the optimal hyperparameters were based on previous reports (He et al., 2016; Hu et al., 2018; Keskar et al., 2016[2]).

To reflect demographic metadata (age and sex), we trained a feed-forward network separately. After calculating the malignancy score using the 178 outputs, these scores were used for the input of the feed-forward network. The feed-forward network consists of three inputs (malignancy score, age, and sex) as an input, three hidden layers with 200 nodes, and the last softmax layer. The feed-forward network was trained using 120 thousand images of the ASAN dataset using the NVIDIA Caffe (https://github.com/nvidia/caffe; version 0.17.2), and the hyperparameters for training was as follows: learning_rate = 0.01, gamma = 0.1, weight_decay = 0.0001, mini_batch_size = 32, solver = SGD, momentum = 0.9, total_iteration = 30 epoch, and step_iteration = 10 epoch.

SUPPLEMENTARY TABLES

Supplementary Tables S1−S8

SUPPLEMENTARY REFERENCES

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 2018a;138:1529−38.

Han SS, Lim W, Kim MS, Park I, Park GH, Chang SE. Interpretation of the outputs of a deep learning model trained with a skin cancer dataset. J Invest Dermatol 2018b;138:2275−7.

Han SS, Moon IJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. JAMA Dermatol 2020a;156:29−37.

Han SS, Park I, Eun 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 2020b;140:1753−61.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 770−778.

Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 IEEE/CVF conference on computer vision and pattern recognition; 2018. p. 7132−7141.

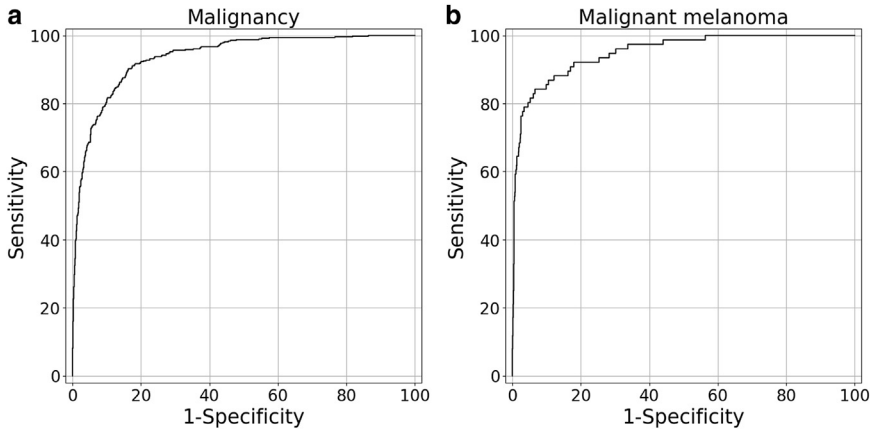
Muñoz-López C, Ramírez-Cornejo C, Marchetti MA, Han SS, Del Barrio-Díaz P, Jaque A, et al. Performance of a deep neural network in teledermatology: a single-centre prospective diagnostic study. J Eur Acad Dermatol Venereol 2021;35:546−53.

Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality? J Invest Dermatol 2018;138:2277−9.

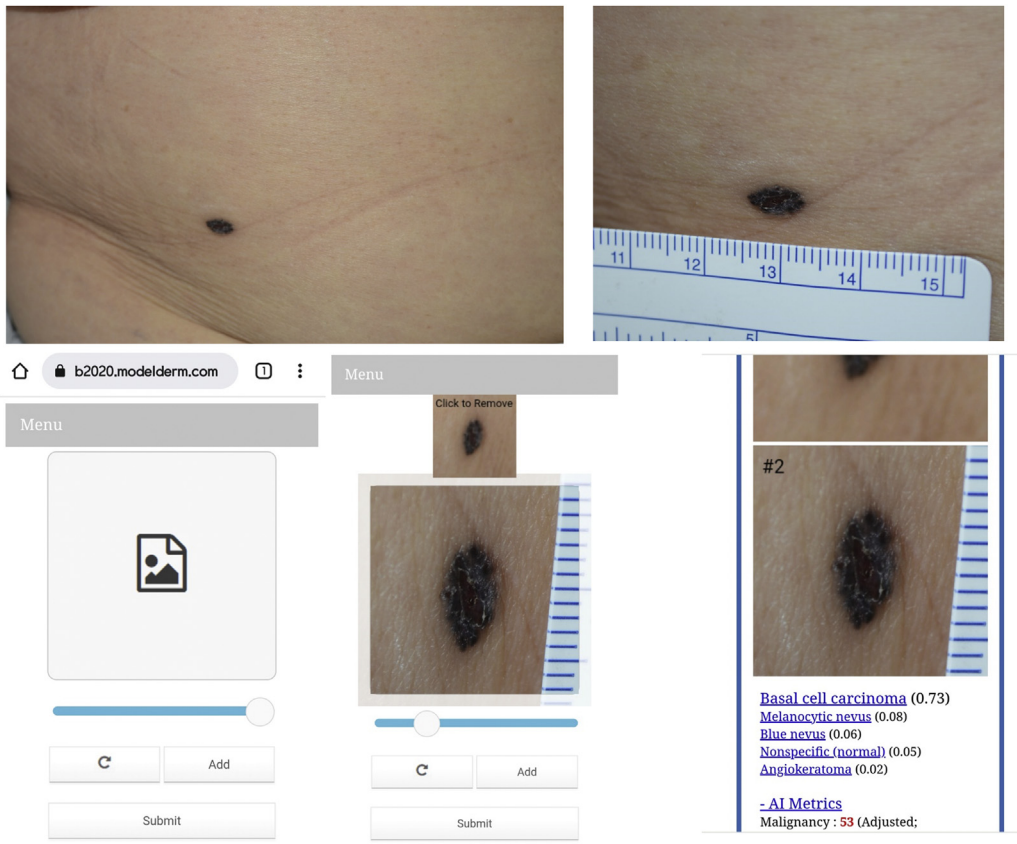
Navarrete-Dechent C, Liopyris K, Marchetti MA. Multiclass artificial intelligence in dermatology: progress but still room for improvement. J Invest Dermatol 2021;141:1325−8.

[1]Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, et al. Resnest: split-attention networks. arXiv 2020.

[2]Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP. On large-batch training for deep learning: generalization gap and sharp minima. arXiv 2016.



Supplementary Figure S1. Binary classification for determining malignancy using the Edinburgh 1,300 images. (**a**) Malignancy determination in the binary classification using the 1,300 images. Area under the curve: 0.937; 95% CI = 0.924−0.950 (DeLong method). (**b**) Melanoma diagnosis in the multiclass classification using the 1,300 images. Area under the curve:0.951; 95% CI = 0.927−0.975 (DeLong method). The ROC curve was drawn using the one-vs-rest methods in the multiclass classification. CI, confidence interval; ROC, receiver operating characteristic.



Supplementary Figure S2. An example using the online algorithm. The clinical photographs were captured in the studio with a brightness of 300 lux either using a softbox or without a flash. The body of the digital camera was either Nikon D7100 or D7500, and the zoom lens was either AF-P DX NIKKOR Zoom 18-55mm f/3.5-5.6G or AF-S DX Nikkor Zoom 18-55mm f/3.5-5.6G. Algorithm's five diagnoses, their probabilities, and malignancy score were used for the experiment. Multiple macro images having the large lesion of interest at the center were uploaded. Interpretation of the Top outputs and malignancy output was instructed as follows: (i) Top output: the Top output range from 0.0 to 1.0 (Top output ≥ 0.2: the predicted diagnosis is a meaningful differential diagnosis and Top output < 0.2: only a small chance for the predicted diagnosis). (ii) Malignancy output: the malignancy output ranges from 0 to 100 (malignancy score ≥ 20: high chance of malignancy; malignancy score ≥10 and <20: still some chance of malignancy; and malignancy score < 10: maybe benign). The patient pictured in this figure consented to the publication of the image.