의학박사 학위논문

학습된 심층신경망의 피부 종양과 조갑 백선에 대한 진단 능력 및 수련의의 진단 능력 향상 기여 평가

Augmented Intelligence in Diagnosing Skin Lesions Suspected of Skin Neoplasms and Onychomycosis

울산대학교대학원

의 학 과

김영재

Augmented Intelligence in Diagnosing Skin Lesions Suspected of Skin Neoplasms and Onychomycosis

지도교수 장성은

이 논문을 의학박사 학위 논문으로 제출함

2022년 2월

울산대학교대학원

의학과

김영재

김영재의 의학박사 학위 논문을 인준함

심사위원 이미우 인

심사위원 원종현 인

심사위원 장성은 인

심사위원 이우진 인

심사위원 오상호 인

울산대학교대학원

2022년 02월

ABSTRACT

Background: Although deep neural networks have shown promising results in the diagnosis of skin cancer and onychomycosis, a prospective evaluation in a real-world setting could confirm these results. This study aimed to evaluate whether an algorithm (http://b2019.modelderm.com, http://nail.modelderm.com) improves the accuracy of non-dermatologists in diagnosing skin neoplasms and onychomycosis.

Methods: A prospective observational study was performed in patients presenting with dystrophic features in the toenails. Five board-certified dermatologists determined a diagnosis of onychomycosis using the clinical photographs. The diagnosis was also made using the algorithm and dermoscopic examination to evaluate the diagnostic abilities of a deep neural network (http://nail.modelderm.com) for onychomycosis. For skin neoplasms, random series cases with skin neoplasms suspected of malignancy by either physicians or patients were recruited in two tertiary care centers located in South Korea. An artificial intelligence (AI) group was diagnosed via routine examination with photographic review and assistance by the algorithm, whereas the control group was diagnosed only via routine examination with a photographic review. The accuracy of the non-dermatologists before and after the interventions was compared. A randomized trial (KCT0005614) was also conducted to validate whether artificial intelligence (AI) could augment the accuracy of non-expert physicians in the real-world setting which included diverse out-of-distribution conditions. Intern doctors and dermatology residents examined the randomly allocated patients with suspicious skin lesions with or without the real-time assistance of AI algorithm (https://b2020.modelderm.com#world). We compared the change in accuracy, sensitivity, and specificity before and after the assistance of the algorithm, to confirm the performance of augmented intelligence.

Results: In onychomycosis study, a total of 90 patients (mean age, 55.3; male, 43.3%) assessed between September 2018 and July 2019 were included. The detection of onychomycosis using the algorithm (AUC, 0.751; 95% CI, 0.646–0.856) and that by dermoscopy (AUC, 0.755; 95% CI, 0.654–0.855) were seen to be comparable (Delong's test; *p* = 0.952). The sensitivity and specificity of the algorithm at the operating point were 70.2% and 72.7%, respectively. The sensitivity and specificity of diagnosis by the five dermatologists were 73.0% and 49.7%, respectively. The Youden index of the algorithm (0.429) was also comparable to that of the dermatologists' diagnosis (0.230 ± 0.176; Wilcoxon rank-sum test; $p = 0.667$).

For skin neoplasms, among the AI group, the accuracy of the first impression (Top-1 accuracy; 58.3%) after the assistance of AI was higher than that before the assistance (46.5%, $p = 0.008$). The number of differential diagnoses of the participants increased from 1.9 ± 0.5 to 2.2 ± 0.6 after the assistance ($p < 0.001$). In the control group, the difference in the Top-1 accuracy between before and after reviewing photographs was not significant (before, 46.1%; after, 51.8%; $p = 0.19$), and the number of differential diagnoses did not significantly increase (before, 2.0 ± 0.4; after, 2.1 ± 0.5; $p = 0.57$).

In randomized controlled study, using 576 consecutive cases with suspicious lesions, the accuracy of the AI group (n = 295, 52.5%) was significantly higher than those of the Unaided (n=281, 43.4%; $p = 0.035$). The augmentation was more significant from 53.3% (n = 150) to 29.7% (n=138; $p < 0.0001$) in the intern doctors who had the least experience in dermatology, whereas the augmentation was minimal in dermatology residents. The algorithm could help the trainees in the AI group consider more differential diagnoses than the Unaided (2.09 versus 1.95; $p = 0.0005$).

Conclusion: As a standalone method, the algorithm analyzed photographs taken by non-physician and showed comparable accuracy for the diagnosis of onychomycosis to that made by experienced dermatologists and by dermoscopic examination. For the diagnosis of skin neoplasms, AI augmented the diagnostic accuracy of trainee doctors in real-world settings. This result was also confirmed in a single-center, unmasked, paralleled, randomized controlled trial.

Keywords: Augmented intelligence, Artificial intelligence, Skin neoplasm, Onychomycosis, Diagnostic accuracy

Contents

LIST OF FIGURES AND TABLES

INTRODUCTION

Artificial intelligence (AI) has demonstrated performance comparable with that of specialists in the medical field.(1) In dermatology, AI could analyze dermoscopic and clinical images as accurately as dermatologists in reader tests.(1-7) However, these studies were all retrospective and mostly reader-tested for selected cases, which have complicated translation to actual practices for several limitations. There is limited data on whether the algorithm's decision can really lead to a change in the clinician's decision in a prospective real-world setting.

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 a prospective study using 340 consecutive teledermatology cases,(8) the Top-1 accuracy of the algorithm (41.2%) was lower than that of the general practitioners (49.3%) because 10.3% of the teledermatology cases belonged to the untrained classes ('out-of-distribution'). When the analysis was limited to the explicitly trained diagnoses ('in-distribution'), the balanced Top-1 accuracy of the algorithm (47.6%) was comparable to the dermatologists (49.7%).

In the field of dermatology, only a small number of prospective studies with diagnostic AI have been reported.(3, 8-12) Moreover, only 12 randomized controlled trials (RCT) were published in year of 2021, and no RCT studies have been published in dermatology. In a prospective study, an algorithm demonstrated the ability to identify melanoma with an accuracy similar to that of specialists.(3) An AI algorithm for the diagnosis of onychomycosis trained with 49,567 nail images overwhelmed 42 dermatologists in a retrospective study.(13) This result was promising because the conventional diagnostic tools for onychomycosis including direct microscopic examination with potassium hydroxide (KOH) and fungal cultures have been complex, time-consuming, and may be distressing for the patient due to the need for scraping.

We have developed a skin disease classifier (Model Dermatology; https:// modelderm.com) to diagnose 178 skin diseases and predict the chance of malignancy in previous studies.(4, 14, 15) Algorithm for the diagnosis of onychomycosis was also developed (http://nail.modelderm.com). In cases with onychomycosis, we performed a prospective, observational comparative study aimed to evaluate the diagnostic power of an algorithm in comparison with diagnosis made by experienced dermatologists and dermoscopic examination. In cases with skin neoplasms, we first performed a prospective study with random series at two tertiary care centers in Korea to demonstrate whether the accuracy, sensitivity, and specificity of trainees improved with the assistance of an algorithm in real-world practice. After confirming the diagnostic accuracy, we finally performed a single-center, paralleled, unmasked, RCT to investigate whether a multiclass AI algorithm could instantly improve the accuracy and sensitivity/specificity of non-dermatologists who examined patients with suspicious skin neoplasms detected by either a patient or physician. In addition, we compared the change in accuracy, sensitivity, and specificity before and after the assistance of the algorithm, to confirm the performance of augmented intelligence.

MATERIALS AND METHODS

Artificial intelligence for the diagnosis of onychomycosis

A prospective, observational comparative study was conducted at a tertiary hospital between September 2018 and July 2019. The study design was approved by the Institutional Review Board of Asan Medical Center (IRB number: 2018-1368).

Patients presenting with a dystrophic toenail were enrolled (Table 1). To perform KOH evaluation, fungal culture, dermoscopic examination, and algorithm analysis in the same nail, targeted toes were identified by a skin marker. Clinical photographs of the whole foot were taken by research assistants. Direct microscopy with KOH 40% and culture were performed to confirm the diagnosis in all cases. The ground truth was determined either by direct microscopy with KOH testing or by fungal culture.

Five board-certified dermatologists (with a mean of 5.6 years of experience) determined a diagnosis of onychomycosis using the clinical photographs. Dermoscopic examination was performed using established diagnostic criteria (16) by two board-certified dermatologists. All dermoscopic features were recorded on a 10-point scale.

In a previous study, (13) we created and released onychomycosis convolutional neural network models (Figure 1); the same algorithm was used in this study without modification (http://nail.modelderm.com). The operating cut-off of the algorithm was obtained using the datasets (342 patients; 780 onychomycosis and 578 nail dystrophy images), which were used as the validation dataset in the previous study.(13) The optimal point that maximizes the sum of sensitivity and specificity was used as the operating cut-off threshold in this study.

Receiver operating characteristic (ROC) curves were drawn using each score of the algorithm and dermoscopic examination. The area under the curve (AUC; pROC package version, 1.15.3; R version 3.4.4) was calculated, and sensitivity, specificity, and Youden index score (sensitivity+specificity-100%) were compared between results of the algorithm, clinician evaluation, and dermoscopic examination. Wilcoxon rank-sum test was used to compare the variables. Delong's test was performed to determine whether two ROC curves were statistically different.

Artificial intelligence for the diagnosis of skin neoplasms: a prospective controlled before-and-after study

After obtaining approval from the institutional review board of Asan Medical Center (2018-1130), a prospective study was performed at two tertiary care centers in Korea (230 cases from Department of Dermatology, Asan Medical Center, and 55 cases from Seoul National University, Bundang Hospital) between February 1, 2020, and November 7, 2020. The algorithm (Model Dermatology, build 2019; https://b2019.modelderm.com) developed in our previous study(4, 17) was used. The algorithm suggests the three most probable diagnosis of uploaded photographs and also reports a malignancy score (range: 0–100).

After obtaining informed consent, all patients (age>19 years) who had skin neoplasms suspected of malignancy by either patient or physician were recruited. Exclusion criteria were patient refusal, broken blindness, the wrong version of the algorithm, non-real-time analysis, and exposure of the biopsy results in the referral note (Supplementary figure 1). If first impressions were recorded at >24 h after patients' visits, they were classified as non-real-time. There were no inconclusive cases in the prediction of the algorithm. Ultimately, 270 pathologically diagnosed cases and 15 clinically diagnosed cases were used in the final analysis (Table 2 and Supplementary table 1). A total of 139 and 131 cases were pathologically diagnosed in the AI group and the control group, respectively. A total of 15 cases (5 cases = AI group, 10 cases = Control group) were clinically diagnosed because the attending physicians concluded that they were definitely benign cases and do not to be biopsied.

A total of 10 attending physicians (11.4 ± 8.8 years' experience after board certification), 11 dermatology trainees, and 7 intern doctors participated in this study (Supplementary table 2). Attending physicians routinely recorded their diagnoses after thorough examinations. The trainees who were blinded to attending physicians' diagnoses evaluated the patients. After quasi-randomization using odd/even patient ID, the trainee took the patient's medical history, performed physical examinations, took photographs, and provided their diagnoses up to three predictions. In the AI group, trainees selected one photograph and uploaded on http://b2019.modelderm.com. After referring to the algorithm's three diagnoses and the malignancy score, they were given an opportunity to modify their initial diagnoses. In the control group, trainees just reviewed the photographs once again then provided the after-diagnoses.

Top accuracy was calculated as an evaluating metric. 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." Only an exact diagnosis was recorded as correct. For evaluating the sensitivity and specificity of malignancy prediction, the physicians' diagnoses were transformed into either malignant or benign. Top accuracies were compared using two-tailed paired Wilcoxon signed-rank tests (R version 3.5.3), and a p-value of <0.05was considered statistically significant.

Artificial intelligence for the diagnosis of skin neoplasms: a randomized controlled trial

This was an Institutional Review Board of Asan Medical Center (S2018-1703-0001) approved prospective study. The study was performed in the Department of Dermatology at Asan Medical Center, a tertiary care center in Seoul, Korea. The study was conducted from November 30, 2020, to September 9, 2021 after the registration (cris.nih.go.kr; KCT0005614). The development of the algorithm (Model Dermatology, Build2020; https://b2020.modelderm.com#world) is described in the Supplementary method, and the algorithm was fixed on Sep 19, 2020. Along with 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 previously used.(17) Using the subset of the SNU 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%.(18) The high-specificity threshold was defined as the threshold at which 80% sensitivity was obtained.

In our prospective before-and-after study, 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, power = 0.8), and we planned to recruit 600 cases.(19)

All patients signed informed consent prior to inclusion in the study. We included adult consecutive patients (age>19 years) who had one or more suspicious skin lesion of skin cancer detected by either patient or physician. Exclusion criteria of patients and input data included patient refusal (10 cases), wrong recruitment (6 cases; age≤19 years), biopsy refusal (2 cases), and non-real-time analysis (9 cases) (Supplementary figure 2). 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 were no performance errors 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 3). A total of 53 conditions were within the trained 178 classes ('in-distribution') and 30 conditions were not trained by the algorithm ('out-of-distribution') (Supplementary table 3).

A total of 4 attending physicians (3, 4, 6, and 22 years of experience after board certification), 4 first-year dermatology residents, and 4 intern doctors (first year after getting a medical license in Korea) participated in this study. Attending physicians routinely recorded their impressions after thorough examinations. After the simple randomization using a custom randomizer 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. The clinical photographs were captured in the main 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.

In the AI group, trainees selected 1~3 photographs with age and gender 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 of the algorithm's diagnoses and malignancy score. The photographs that the trainees judged to be of adequate quality were uploaded by the trainees. In the Unaided group, trainees examined routinely and recorded the three most probable diagnoses, without the assistance of the algorithm. The use of dermoscopy was not allowed for all trainees.

In calculating Top accuracy, only an exact diagnosis was recorded as correct, but the subtype of the disease was counted to be correct. For example, intradermal nevus was counted correct for the ground truth of junctional nevus. We manually lumped together 364 diagnoses in natural language into the 83 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-squared test with Yates' continuity correction (AI group versus Unaided group) or McNemar's test (Before versus After the assistance of the algorithm in the AI group) using R version 4.1.1, and a p-value of <0.05 was statistically significant.

RESULTS

Artificial intelligence for the diagnosis of onychomycosis

A total of 90 patients (mean age, 55.30 ± 14.13 years; male, 44.3%) were included in the study (Table 1). KOH positivity was 84.2% (n = 48), culture positivity was 54.4% (n = 31), and positivity for both KOH and culture was 24.4% (n = 22). Since the ground truth was determined by either direct microscopy with KOH test or fungal culture, 63.3% of patients (n = 57) were diagnosed with onychomycosis.

The AUC value of the algorithm was 0.751 (95% CI, 0.646–0.856), and the sensitivity / specificity of the algorithm at the cut-off threshold were 70.2 / 72.7% (Figure 2 and 3). The AUC value of dermoscopic examination was 0.755 (95% CI, 0.654–0.855), and the sensitivity / specificity at the optimal operating point of the dermoscopic examination were 72.7 / 72.9%, respectively. Delong's test showed no significant difference between the ROC curves of the algorithm and dermoscopic diagnosis (p = 0.952).

The mean sensitivity and specificity of diagnosis by five board-certified dermatologists were 73.0% ± 14.7% and 49.7% ± 7.6%, respectively. The mean Youden index of the five board-certified dermatologists was 0.230 ± 0.176, which was comparable to that of algorithm (0.429) using Wilcoxon rank-sum test (p = 0.667).

The positive predictive value / negative predictive value of the algorithm were 73.4% (95% CI, 61.5–82.7) / 61.5% (95% CI, 35.5–82.3), and those of dermoscopic examination were 69.3% (95% CI, 58.2–78.6) / 66.7% (95% CI, 41.7–84.8), and those of the five dermatologists were 76.8% ± 8.4% and 56.9% ± 15.5%, respectively.

Artificial intelligence for the diagnosis of skin neoplasms: a prospective controlled before-and-after study

Result of the AI Group

After analyzing the accuracies before and after assistance, it was noted that the Top-1 / Top-2 / Top-3 accuracies after assistance were significantly higher than those before assistance (before = 46.5% / 54.2% / 54.9%; after = 58.3% / 70.1% / 71.5%; p = 0.008 / <.001 / <.001) (Figure 4). The Top-1 / Top-2 / Top-3 accuracies of the attending dermatologists were 61.8% / 69.4% / 71.5%, respectively, and those of the standalone algorithm were 53.5% / 66.0% /

70.8%, respectively. In 42.4% (61 / 144) cases, the Top-1 diagnosis of the algorithm was coherent with that of the trainees, and in 50.0% (72 / 144) cases, the Top-1 of the algorithm was coherent with that of the attending physicians. The Top-1 of the trainees was coherent with that of the attending physicians in 52.8% (76 / 144) cases.

The trainees revised 28.5% (41 / 144) of their Top-1 diagnosis after reviewing three diagnoses of the algorithm. A total of 70% (29 / 41) of their revised answers were correct, whereas 29% (12 / 41) of their revised answers were incorrect.

For determining malignancy, the sensitivity / specificity derived from the Top-1 was 78.3% / 88.4% before the assistance and 73.9% / 94.2% after the assistance (Table 4, $p = 0.77$ / = 0.06). The sensitivity / specificity of the attending dermatologists was 82.6% / 91.7% and that of the patients were 56.5% / 42.6%. The sensitivity / specificity derived from the Top-1 diagnosis of the algorithm was 52.2% / 93.4%. The sensitivity / specificity at the threshold of the risk “Medium” using the malignancy score was 95.7% / 60.3% and that at the threshold of the risk “High” was 82.6% / 70.2% (Table 4). The number of differential diagnoses by the trainees increased from 1.9 ± 0.5 to 2.2 ± 0.6 ($p < 0.001$).

Result of the Control Group

The differences of the Top-1 / Top-2 / Top-3 accuracies between before and after reviewing photographs were not significant (Control-Before, 46.1% / 64.5% / 66.7%; Control-After, 51.8% / 66.7 / 68.1%; $p = 0.19$ / = 0.42 / = 0.35).

For determining malignancy, the sensitivity / specificity derived from the Top-1 diagnosis was 65.5% / 81.3% before reviewing and 65.5% / 86.6% after reviewing (Table 4, $p = 1.00$ / = 0.09). The sensitivity / specificity of the attending dermatologists was 79.3% / 90.2% and that of the patients was 48.1% / 44.5%.

The number of differential diagnoses by the trainees had not changed significantly (Control-Before = 2.0 ± 0.4, Control-After = 2.1 ± 0.5; $p = 0.57$).

AI Group versus Control Group

The differences of the Top-1 / Top-2 / Top-3 accuracies between the AI group and the Control were not significant (AI Group = 58.3% / 70.1% / 71.5%; Control Group = 51.8% / 66.7% /

68.1%; p = 0.27 / = 0.53 / = 0.53). Summarized key results were described in Supplementary table 4.

Artificial intelligence for the diagnosis of skin neoplasms: a randomized controlled trial

AI group versus Unaided group

To confirm that the two groups were truly comparable, the accuracies of the attending dermatologists and trainees (before interventions) were compared. The Top-1 accuracy of attending dermatologists (62.3%) and trainees (43.4%) of the Unaided group were higher than those of the AI group (dermatologists = 59.3%, trainees = 40.0%), which indicated that easier cases were not allocated to the AI group (Table 5).

The Top-1 accuracy of the AI group was 52.5% and that of the Unaided was 43.4% (p = 0.035; Figure 5, Table 5, Supplementary table 5). There were significant differences in the result depending on whether the participant was an intern or a dermatology resident. The Top-1 accuracy of the AI-intern (53.3%) was markedly higher than that of the Unaided-intern (29.7%; p < 0.0001) whereas the Top-1 accuracy of the AI-resident and Unaided-resident was 51.7% and 56.6%, respectively (p = 0.47). In the AI group, we compared the judgment before and after receiving the assistance of the algorithm, and there was a significant enhancement in the Top-1 accuracy of the AI-intern group (interns = 30.0%, augmented interns = 53.3%, p < 0.0001; Supplementary table 6). However, in the AI-resident group, the change was minimal (residents = 50.3%, augmented residents = 51.7%; p = 0.86) As shown in Supplementary table 7, a larger improvement in accuracy was observed for the subsequent cases, up to +22.0% / +31.7% for the Top-1 / 3 accuracy, although the accuracy for Top-1 / 3 was improved by only +0.0% / +14.0% for the first 10 cases.

When the analysis was restricted within 266 cases that were biopsied, the Top-1 / 3 accuracy of the standalone algorithm, trainees, augmented trainees, and attending dermatologists were 39.5% / 48.1%, 54.5% / 69.5%, and 54.9% / 64.7%, 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 / 3 accuracy of trainees and attending dermatologists was 42.2% / 56.6% and 58.9% / 68.2%, respectively.

Malignancy determination affects clinical decisions such as performing 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 was 84.2% / 69.3% and 75.6% / 63.1%, respectively (p = 0.48 / 0.18; Table 6). The sensitivity / specificity of the AI-intern group and the Unaided-intern was 80.0% / 81.5% and 56.3% / 68.9%, respectively (p = 0.24 / 0.029). The sensitivity / specificity of the AI-resident group and the Unaided-resident was 88.9% / 56.7% and 86.2% / 57.0%, respectively (p = 1.0 / 1.0).

There was a significant difference in the number of differential diagnoses between the AI group and the Unaided (2.09 versus 1.95; Wilcoxon rank-sum test, p = 0.0005). The diagnosis number of the AI-intern group (2.00) was higher than the Unaided-intern (1.88; Wilcoxon rank-sum test, p < 0.0001). The diagnosis number of the AI-resident group (2.17) was also higher than that of the Unaided-resident (2.01; Wilcoxon rank-sum test, p = 0.019).

Before and After Comparison – Individual Analysis

Individual improvement in Top-1 diagnostic accuracy for each trainee ranged from -5.3% to +41.4%, with an average of +12.4%, which was not statistically significant (paired t-test with Shapiro test; p = 0.059; Supplementary table 8). On the other hand, Top-3 accuracy was improved by +0% ~ +41.4%, with an average improvement of +20.8%, which was statistically significant (paired t-test with Shapiro test; p = 0.0025). There was individual variation in the degree of improvement: the diagnostic accuracy of one resident did not change at all for both Top-1 and Top-3 (Top-1 = -5.3%, Top-3 = +0.0%) whereas that of one intern was improved by +41.4% for both Top-1 and Top-3 (Supplementary table 8).

Standalone Performance of the Algorithm

The standalone Top-1 / 3 accuracy of the algorithm in the AI Group was 49.2% / 72.9%. The AUC for determining malignancy was 0.889 (95% CI 0.831–0.947; DeLong method), which was equivalent to that of the attending dermatologists on the ROC curve (Figure 6). At the high-sensitive threshold, the sensitivity / specificity was 92.1% / 61.9%, and at the high-specificity threshold, the sensitivity/specificity was 84.2% / 78.2%. The sensitivity / specificity of the standalone algorithm derived from the Top-1 was 65.8% / 90.3% and that from the Top-3 was 84.2% / 61.1% (Table 6).

DISCUSSION

Onychomycosis is a common nail disorder accounts for approximately 40% of all nail disorders.(20) Despite its high prevalence and clinical importance, it is challenging for clinicians to diagnose onychomycosis due to its similarity to other nail disorders. Traditionally, mycological diagnosis was made using KOH examination or fungal cultures. The sensitivity and specificity of these tests were estimated to be 52.5–81.8% and 72.0–100%, respectively for KOH, and 57.0–59.0% and 82.0–100%, respectively, for fungal culture.(20-22) However, the two tests require the use of specific equipment and are time-consuming, particularly culture, which requires at least 4 weeks' incubation. New diagnostic tools involving histopathologic examination using Periodic acid-Schiff staining of nail clippings have shown greater sensitivity (88.2–93.1%) but cannot provide an immediate diagnosis in the clinical setting.(23) The algorithm used in this study demonstrated comparable accuracy to the diagnosis of dermoscopic features. Unlike KOH and dermoscopic examination, which are time-consuming and must be carried out by well-trained personnel, diagnosis using artificial intelligence can be made using photographs taken by non-physicians in a real-time setting.

Unlike previous studies, our study is designed particularly for assisting non-dermatologists rather than dermatologic experts, and the algorithm is fully opened and accessible through the website. This aspect of our algorithm enables patients to screen their onychomycosis on a daily life without the help of the specialists. When we analyzed the area involvement of nail, 65.0% of patients revealed nail involvement in less than half of total nail area (Table 1). Relatively higher frequency of mild cases in this study implies more beneficial value of our algorithm in patients' daily self-practical application.

In cases with skin neoplasms, we found that the AI assistance improved the diagnostic accuracy of trainee doctors in a prospective before-and after study. Owing to various biases, the outstanding performance of algorithms may not always be reproduced in real-world settings.(24, 25) Because algorithms cannot be trained for all diseases, they may show false positives for various out-of-distributed conditions. Both the metadata and photographs used in training and reader testing could be biased if handled by different expertise. For example, dermatologists may take few photographs of nail hematoma because they diagnose it with full confidence, and the algorithm trained with a few cases of hematoma may show uncertainty.

Therefore, clinical validation should be performed with the same level of expertise as the end-user.

To date, the incorporations of AI into dermatological practice have been steadily investigated.(1-7) It was revealed that a trained classifier algorithm could execute diagnostic performance as equal as dermatologists for clinical and dermoscopic images of suspected melanoma and carcinoma.(1) Haenssle et al.(26) demonstrated that AI could correctly classify dermoscopic images of suspected melanoma into benign, in situ, or invasive at levels equal to and greater than expert dermatologists. Another recent study found that the performance of AI trained with dermoscopic images for identifying melanoma showed dermatologist-level image classification on a clinical image classification task. The mean sensitivity and specificity achieved by the 145 dermatologists with clinical images was 89.4% and 64.4%, whereas AI showed a mean specificity of 68.2% at the same sensitivity.(2)

In our previous study, we also found that trained AI could classify clinical images into 12 common cutaneous diseases including skin neoplasms (basal cell carcinoma, squamous cell carcinoma, intraepithelial carcinoma, actinic keratosis, seborrheic keratosis, malignant melanoma, melanocytic nevus, lentigo, pyogenic granuloma, hemangioma, dermatofibroma, and wart) with similar sensitivity and specificity of dermatologists.(4)

Reflecting these points on the diagnostic excellence of AI, the concept of augmented intelligence has recently emerged. Augmented intelligence is a term that focuses on the assistive role of AI, emphasizing that augmented intelligence is designed to enhance human intelligence and the clinician-patient relationship rather than substitute it.(27) The American medical association (AMA) states that augmented intelligence algorithms should be clinically validated before being integrated into patient care.(28) Therefore, they strongly recommended performing prospective clinical trials evaluating safety and effectiveness with relevant clinical end points. Despite these recommendations, previous studies incorporating AI into dermatological practice have not been prospectively verified in the real-world setting.

In this study, although the Top-1 accuracy of the standalone algorithm (53.5%) was comparable with that of the trainees (46.5%), the Top-1 accuracy of the augmented trainees (58.3%) was significantly higher. This augmentation could be owing to different strategies between humans and AI algorithm.(29, 30) The coherence between the algorithm–human (algorithm–trainees = 42.4%; algorithm–attending dermatologists = 50.0%) was lower than that between human–human (trainees–attending dermatologists = 52.8%), which implied different diagnostic patterns.

The augmentation may be achieved when the accuracy of the algorithm is higher or at least comparable with that of the user. In the study using dermoscopic images, the physicians with the least experience were the most frequently augmented.(31) For neoplastic skin lesions, the diagnostic accuracy of non-dermatologists has been reported to be 40%–47%.(32) In this study, the Top-1 accuracy of the trainees improved from 46.5% to 58.3% (25.4% increase) instantly by referring to the second opinion of the algorithm.

The sensitivity derived from the Top-1 prediction of the algorithm was low (52.2%), as noted previously.(12) Consequently, the sensitivity of the trainees derived from the Top-1 may decrease from 78.3% to 73.9% ($p = 0.76$). Our algorithm was developed with numerous benign crops to cope with the false-positive problem in detecting skin cancer using unprocessed images(17) and a multitude of benign crops in the training dataset could distort the overall output trend, making it more likely to predict benign conditions.

We further performed randomized controlled trial to investigate the augmentation of AI in the diagnosis of skin neoplasms. We demonstrated that a multi-class AI algorithm helped to improve the diagnostic accuracy and specificity of the trainees. The augmentation was significant in the intern doctors who had the least experience in dermatology, whereas the augmentation was minimal in dermatology residents. Regarding the standalone performance with 266 biopsied cases, the accuracies of the AI-augmented trainees were comparable with those of the attending dermatologists. In addition, the standalone algorithm using the malignancy score demonstrated comparable performance with attending dermatologists in determining malignancy. This is a unique result because this study was conducted in the real-world setting which included diverse out-of-distribution conditions.

Although several retrospective studies have demonstrated successful results on the diagnosis of skin lesions using AI algorithms, studies were carried out in experimental settings,(33) and various factors make these promising results not to be reproduced in real-life. First, Clever-Hans type bias may affect the results.(34) 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. The second factor is 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 performance with the in-distribution 134 disorders,(4) the performance deteriorated in the prospective study(8) with consecutive patients having diverse disorders, which indicates the relevance of the out-of-distribution problem. Even if algorithms are trained on rare diseases, the diagnostic ability may be poor because the small amount of training data for these ‘rare’ disease is still not sufficient to train algorithm. Third, the presence of ‘out-of-distribution’ in characteristics may affect the diagnostic outcome. In retrospective experiments, cases with typical features are selected while 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. Fourth, disease prevalence of training dataset may affect the diagnostic accuracy. 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. An algorithm may be simply trained on the disease prevalence of the training dataset, rather than learning the disease features. Finally, there is an unpaired comparison problem between AI and clinicians.(35) Dermatologists do not diagnose relying solely on photographs. The clinicians in the real world use all clinical inputs (i.e., history, touch, body distribution among others). In most circumstances, history taking and physical examination significantly improve the physician's diagnostic ability.

Even if algorithms outperformed dermatologists in previous retrospective studies, the algorithms may perform equivalent or demonstrate lower performance than the dermatologists in real-world prospective studies. In this study, suspected skin neoplasms are 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,(4, 18) and most kinds of tumorous disorders were in-distribution. Nevertheless, in this study, the standalone Top-1 accuracy of the algorithm (49.2%) was inferior to that of the attending dermatologists (59.3%) in the real-world setting.

Even though 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%).(18) In that study,(18) the sensitivity of the dermatologists in the reader test was 84.9%, which was comparable to that of the dermatology residents in the real-world setting of this study (Unaided group=86.8%; AI group=85.8%). The importance of in-person examination was also shown in a study including dermoscopic images in which the diagnostic accuracy of the reader test was lower than that of the physicians who actually performed the dermoscopic evaluation.(36)

There was a marked improvement in the accuracy of the intern doctors who have the least experience in dermatology as previously reported.(5) Another interesting finding is that the augmented accuracy of the trainees (Top-1 / 3 = 54.5% / 69.5%) was equivalent to attending physicians (Top-1 / 3 = 54.9% / 64.7%) for the biopsied 266 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. The multiple 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 huge amount of data. It may be better for humans to understand the diagnostic strengths and limitations of the AI algorithms, and to adapt to the diagnostic characteristics of the machines. The malignancy score of the algorithm on the ROC curve showed the equivalent performance to that of attending dermatologists for determining malignancy on trainees (Figure 6). At the high-sensitivity cut-off threshold, the malignancy score showed 92.1% sensitivity that can compensate for the low sensitivity (63.2%) derived from the Top-1 prediction. However, trainees did not demonstrate synergy in the binary determination as much as they did in the augmented accuracy. In addition, there was no increase in Top-1 accuracy in the first 10 cases, but after that, there was an increase, which means that it might took time for the participants to adapt and use the algorithm (Supplementary table 7). If the participants had a better understanding of the characteristics of the algorithm, the results may be further improved.

Therefore, detailed instructions on the diagnostic characteristics of algorithms should be provided for the users to improve diagnostic accuracy.

Limitation

The algorithm for the diagnosis of onychomycosis used here has several limitations. First, because this study was performed in a tertiary hospital, results with the cases in primary center should be further investigated in multicenter large studies. Second, the results can be significantly affected by the quality of the input images.(13) This has been demonstrated in the previous study, where poor-quality photographs were associated with less accurate diagnostic capabilities.(13) Failed cropping may occur if the photographs obtained by non-physicians are inadequate (Figure 3). Although an ancillary algorithm that can exclude inadequate photographs can accommodate this problem, the impact of image quality on diagnostic accuracy should be further assessed. Lastly, diagnostic approaches in a real practice setting should be processed after checking the clinical features of soles, all toenails, and past medical history.

The algorithm for the diagnosis of skin neoplasms also bears some limitations. Considering that our study population was limited to Asians, our results cannot be generalized in other circumstances. In completely different settings (Asian versus various races, tertiary care versus teledermatology), the standalone accuracy of our algorithm was slightly lower than that of general physicians, although the algorithm could help increase the confidence of the dermatologists.(8) Because the prediction of the algorithm greatly relies on the characteristics of the training data, it may exhibit uncertainty in different settings. Deep learning-based algorithms reflect morphological features and even disease prevalence of the trained dataset; thus, algorithms show the best performance in the same environment.

In before-and after study, patients were randomly recruited but were not recruited consecutively. Therefore, the two groups were not truly comparable. As shown in Supplementary table 1, the cases of BCC and SCC in situ were not assigned evenly, and as shown in Supplementary table 2, the intern doctors with the least experience were more assigned to the AI Group. In randomized controlled study, there may be a hidden bias such as Clever-Hans type(34)) that we were not aware of because the clinical images of Asan Medical

Center were part of the training.

In addition, validation was conducted only on Asians, most with skin types 3 and 4. To enable generalizability, further prospective studies should be performed because disease prevalence, subtype distribution, and visual characteristics of disorders may differ between countries and regions. The retrospective result of the algorithm using the Edinburgh dataset of white population (1,300 images; Top-1 / 3 accuracy = 65.2% / 84.8%, AUC for determining malignancy = 0.937; Supplementary fig 3, Supplementary table 9) should be validated in the further prospective studies. Lastly, only 8 melanoma cases were recruited in this study. Because melanoma prevalence is relatively low in skin types 3 and 4, further study is warranted including other skin types.

REFERENCE

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.

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

3. 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(10):e1913436.

4. Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, et al. Augmented Intelligence Dermatology: Deep Neural Networks Empower Medical Professionals in Diagnosing Skin Cancer and Predicting Treatment Options for 134 Skin Disorders. J Invest Dermatol. 2020;140(9):1753-61.

5. 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(8):1229-34.

6. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020;26(6):900-8.

7. 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(1):137-43.

8. 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(2):546-53.

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.

10. Kim YJ, Han SS, Yang HJ, Chang SE. Prospective, comparative evaluation of a deep neural network and dermoscopy in the diagnosis of onychomycosis. PLoS One. 2020;15(6):e0234334.

11. 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(2):353-9.

12. Navarrete-Dechent C, Liopyris K, Marchetti MA. Multiclass Artificial Intelligence in Dermatology: Progress but Still Room for Improvement. J Invest Dermatol. 2021;141(5):1325-8.

13. 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(1):e0191493.

14. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. J Invest Dermatol. 2018;138(7):1529-38.

15. 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. 2018;138(10):2275-7.

16. Ramos Pinheiro R, Dias Domingues T, Sousa V, Galhardas C, Apetato M, Lencastre A. A comparative study of onychomycosis and traumatic toenail onychodystrophy dermoscopic patterns. J Eur Acad Dermatol Venereol. 2019;33(4):786-92.

17. 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. 2020;156(1):29-37.

18. 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. 2020;17(11):e1003381.

19. Rosner B. Fundamentals of Biostatistics The 7th edition ed. Boston, MA: Brooks/Cole; 2011.

20. Venkateswaramma Begari PP, Anant A. Takalkar. Comparative evaluation of KOH mount, fungal culture and PAS staining in onychomycosis. Int J Res Dermatol. 2019 Aug;5(3):554-8.

21. Weinberg JM, Koestenblatt EK, Tutrone WD, Tishler HR, Najarian L. Comparison of diagnostic methods in the evaluation of onychomycosis. J Am Acad Dermatol. 2003;49(2):193-7.

22. Nada EEA, El Taieb MA, El-Feky MA, Ibrahim HM, Hegazy EM, Mohamed AE, et al. Diagnosis of onychomycosis clinically by nail dermoscopy versus microbiological diagnosis. Arch Dermatol Res. 2020;312(3):207-12.

23. Jung MY, Shim JH, Lee JH, Lee JH, Yang JM, Lee DY, et al. Comparison of diagnostic methods for onychomycosis, and proposal of a diagnostic algorithm. Clin Exp Dermatol. 2015;40(5):479-84.

24. Dreiseitl S, Binder M, Hable K, Kittler H. Computer versus human diagnosis of melanoma: evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. Melanoma Res. 2009;19(3):180-4.

25. Han SS, Moon IJ, Na J-I, Kim MS, Park GH, Kim SH, et al. Retrospective Assessment of Deep Neural Networks for Skin Tumor Diagnosis. medRxiv. 2020:2019.12.12.19014647.

26. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018;29(8):1836-42.

27. Kovarik C, Lee I, Ko J. Commentary: Position statement on augmented intelligence (AuI). J Am Acad Dermatol. 2019;81(4):998-1000.

28. American Medical Association. Augmented intelligence in health care [content derived from Augmented Intelligence (AI) in Health Care (Annual Meeting 2018)]. 2018 June [cited 2019 May 25]. In: American Medical Association Homepage [Internet]. Available from: https://www.ama-assn.org/amaone/augmented-intelligence-ai. [

29. Dodge S, Karam L. A study and comparison of human and deep learning recognition performance under visual distortions. In: 2017 26th International Conference on Computer Communication and Networks (ICCCN) [Internet]. IEEE; 2017. p. 1-7. Available from: https://doi.org/10.1109/ICCCN.2017.8038465 [

30. Geirhos R, Meding K, Wichmann FA. Beyond accuracy: quantifying trial-by-trial

behaviour of CNNs and humans by measuring error consistency. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems 33 (NeurIPS 2020) [Internet]. Curran Associates, Inc.; 2020. p. 13890-13902. Available from: https://papers.nips.cc/paper/2020/hash/9f6992966d4c363ea0162a056cb45fe5-Abstract.html

[

31. 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.

32. Sellheyer K, Bergfeld WF. A retrospective biopsy study of the clinical diagnostic accuracy of common skin diseases by different specialties compared with dermatology. J Am Acad Dermatol. 2005;52(5):823-30.

33. 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.

34. 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(1):1096.

35. Genin K, Grote T. Randomized Controlled Trials in Medical AI: A Methodological Critique. Philosophy of Medicine. 2021;2(1).

36. 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(12):Cd011902.

국문요약

배경: 학습된 심층 신경망이 피부종양 및 조갑 백선의 진단에 유망한 결과를 보여주었지만, 실제 임상에서 심층 신경망의 진단 결과가 얼마나 정확한지에 대한 평가가 필요하다. 본 연구는 심층 신경망 알고리즘(http://b2019.modelderm.com, http://nail.modelderm.com)이 피부 질환(피부 종양 및 조갑 백선 의심 병변)의 진단 능력(민감도, 특이도) 및 수련의(인턴 의사와 피부과 전공의)의 진단 능력을 향상시키는 데에 도움되는 정도를 전향적, 비교적 연구를 통해 평가하는 것을 목적으로 하였다.

재료 및 방법: 조갑 백선의 진단 평가를 위해 2018년 9월부터 2019년 7월까지 발톱의 변형이 동반된 환자를 대상으로 서울아산병원에서 전향적 관찰 연구가 수행되었다. 5명의 피부과 전문의가 임상 사진으로 조갑 백선의 진단을 결정하였다. 비교를 위해 심층 신경망 알고리즘(http://nail.modelderm.com)과 피부 확대경 검사를 이용한 진단도 이루어졌다. 한편, 피부 종양 진단 평가를 위해 의사 또는 환자가 악성으로 의심하는 피부 병변이 있는 환자를 무작위 전향적 시리즈로 국내의 두 개의 3차 의료기관(서울아산병원, 분당서울대학교 병원)에서 2020년 2월부터 2020년 11월까지 모집하였다. 인공 지능(AI) 그룹은 1차 진단 이후 임상 사진의 리뷰와 알고리즘의 진단 결과를 참고하여 2차 진단을 시행하였다. 대조군 그룹은 임상 사진의 리뷰만으로 2차 진단을 시행하였다. 인공지능 알고리즘의 중재 전후로 수련의의 진단 정확도를 비교 평가하였다. 확인된 결과를 바탕으로, 무작위 배정 및 다양한 분포 외 조건(out of distribution)을 포함하는 실제 환경에서의 인공 지능(AI)의 진단 증강 능력을 검증하기 위해 서울아산병원에서 2020년 11월부터 2021년 9월까지 무작위 대조 시험(KCT0005614)을 수행하였다. 수련의가 실시간으로 인공지능 알고리즘(https://b2020.modelderm.com#world)의 도움을 받거나, 받지 않는 방식으로 피부 종양 의심 환자를 무작위 배정해 진단하였다. 마찬가지로 인공지능 알고리즘의 결과 참고 전후의 정확도, 민감도, 특이도의 변화를 비교하였다.

결과: 조갑 백선 연구에서는 총 90명의 환자(평균 연령, 55.3세; 남성, 43.3%)가 포함되었다. 알고리즘(AUC, 0.751; 95% CI, 0.646–0.856)을 사용한 조갑 백선의 진단과 피부 확대경 검사(AUC, 0.755; 95% CI, 0.654–0.855)를 사용한 진단은 유사한 진단 검출 능력을 보였다(p = 0.952). ROC(Receiver Operating Characteristic) curve

의 operating point에서 알고리즘의 민감도와 특이도는 각각 70.2%와 72.7%였다. 5명의 피부과 전문의의 진단 민감도와 특이도는 각각 73.0%와 49.7%였다. 알고리즘의 Youden 지수(0.429)도 피부과 전문의의 진단 지수(0.230±0.176; $p = 0.667$)와 유사했다.

피부 종양의 진단 연구의 경우, AI 그룹에서 최우선 진단의 정확도(Top-1 정확도, 58.3%)가 AI 결과 참고 전보다 높았다(46.5%, $p = 0.008$). 참고 후 수련의의 감별 진단 가짓수는 1.9 ± 0.5개에서 2.2 ± 0.6개로 증가하였다($p < 0.001$). 대조군에서는 임상 사진 리뷰 전후의 Top-1 진단 정확도의 차이가 유의하지 않았고(전 46.1%, 후 51.8%, $p = 0.19$), 감별 진단 가짓수도 크게 증가하지 않았다(전 2.0 ± 0.4, 후 2.1 ± 0.5, $p = 0.57$). 576개의 연속 사례를 포함한 무작위 대조 연구에서는 AI 그룹의 진단 정확도(n = 295, 52.5%)가 AI 결과를 참고하지 않은 그룹의 정확도(n = 281, 43.4%, $p = 0.035$)보다 유의하게 높았다. 피부과 진단 경험이 가장 적은 인턴 의사의 경우 진단 능력의 증강이 53.3%(n = 150)에서 29.7%(n = 138, $p < 0.0001$)로 더 유의미한 반면, 피부과 전공의에서는 증강 정도가 유의하지 않았다. 또한 AI 그룹의 수련의가 참고 하지 않은 수련의에 비해 (2.09 대 1.95, $p = 0.0005$)로 보다 더 많은 감별 진단을 고려하는 것으로 나타났다.

결론: 조갑 백선의 진단에 있어 학습된 심층신경망은 임상 사진을 분석하여 숙련된 피부과 전문의와 피부 확대경 검사를 통한 진단에 필적하는 진단 정확도를 보였다. 피부 종양 진단을 위해 인공 지능은 실제 임상 환경에서 수련의의 진단 정확도를 높이고, 고려하는 감별 진단의 수를 유의미하게 증가시켰다. 이 결과는 단일 기관의 무작위 대조 시험에서도 확인되었다.

중요 단어: Augmented intelligence, Artificial intelligence, Skin neoplasm, Onychomycosis, Diagnostic accuracy

FIGURES AND FIGURE LEGENDS

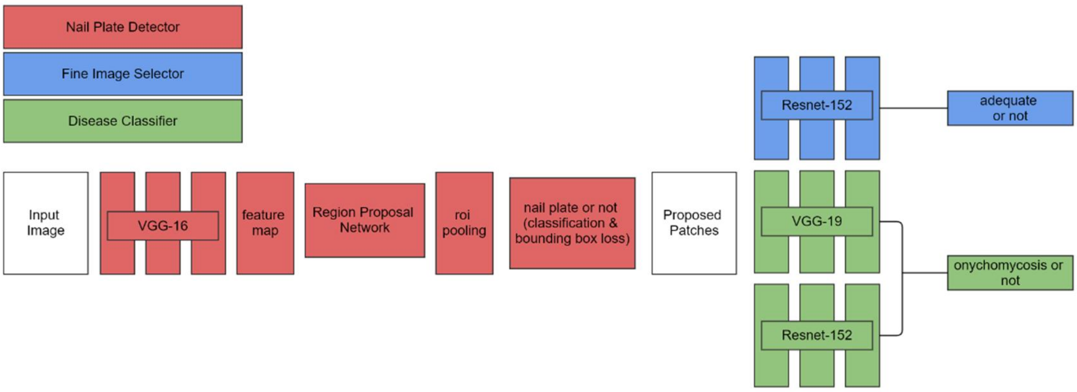


Figure 1. Architecture of the algorithm used in the diagnosis of onychomycosis

Our algorithm comprised three parts: 1) the nail plate detector which detects nail plate from an unprocessed input image, 2) the fine image selector which excludes nail plate images with inadequate quality, and 3) the disease classifier which predicts the chance of onychomycosis. Using the Berkeley Vision and Learning Center (BVLC) deep learning framework Caffe, we fine-tuned the ImageNet pretrained models of ResNet-152 and VGG-19 for the onychomycosis classifier. We also fine-tuned the pretrained model of ResNet-152 for the fine image selector. For the nail plate detector, we used faster-RCNN (backbone network = VGG-16).

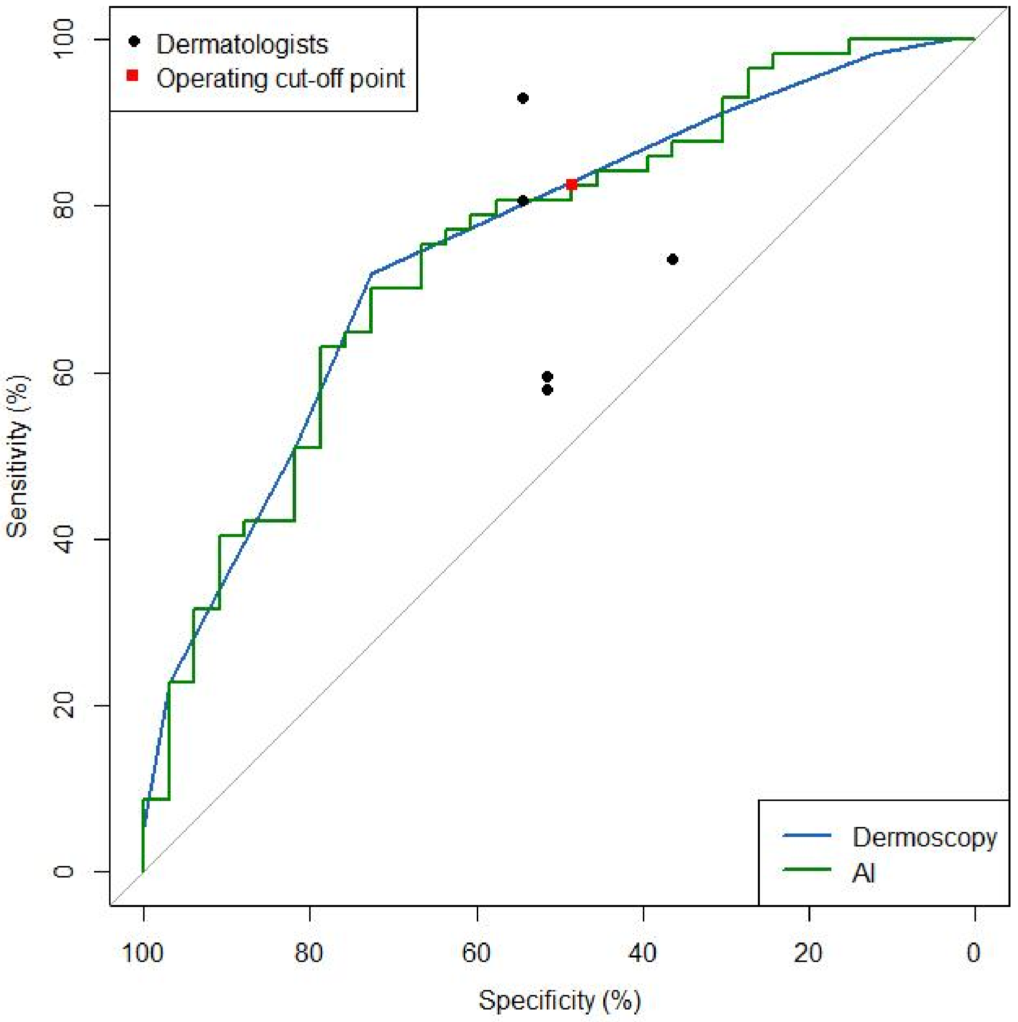


Figure 2. Receiver operating characteristic curves of the algorithm and the dermoscopic examination

The area under the curve (AUC) value of the algorithm was 0.751 (95% CI, 0.646–0.856), whereas the AUC value of dermoscopic examination was 0.755 (95% CI, 0.654–0.855). The results of the reader test are shown as circles (board-certified dermatologists).

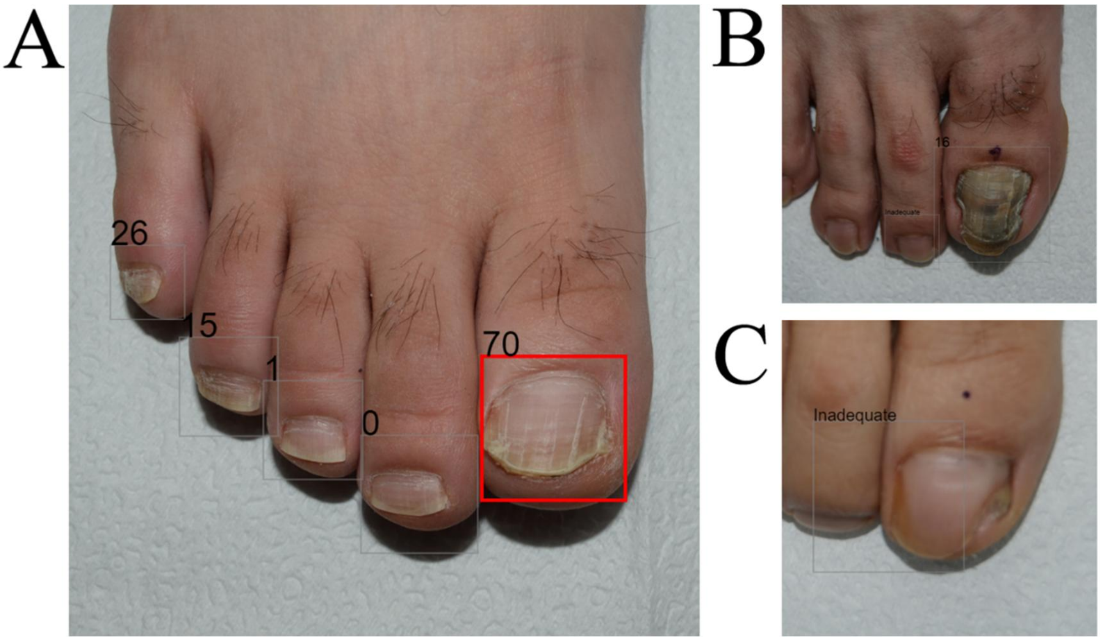


Figure 3. Examples of diagnostic output images in the onychomycosis study

(A) Correct example; a 24-year-old male, confirmed as having onychomycosis by KOH examination and culture. AI made an accurate diagnosis of onychomycosis using this image, whereas two of the five dermatologists misdiagnosed the case as onychodystrophy in the reader test. The rectangle was colored when the onychomycosis output was higher than the operating cut-off threshold (29.3; range 0 – 100).

(B) Incorrect example; a 26-year-old male, confirmed as having onychomycosis by the KOH examination. AI made an inaccurate diagnosis of onychodystrophy using this image, whereas all five dermatologists correctly diagnosed the condition as onychomycosis.

(C) Inadequate quality image; a 49-year-old female, confirmed as having onychomycosis by both KOH examination and culture study. AI first recognized the nail plate, and then the onychomycosis classifier determined whether the nail plate image was onychomycosis or not. With the low-quality, unfocused nail image, AI could not recognize the features properly, resulting in an unreliable diagnostic prediction.

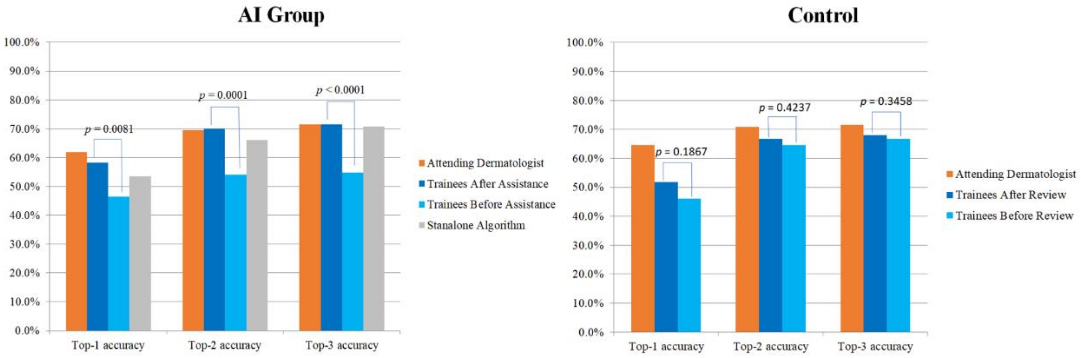


Figure 4. Top accuracies for diagnosing exact diseases in the prospective before-and-after study

The physicians of the AI group (n = 144) referred to the three predictions of the algorithm's diagnoses and the malignancy score before modifying their first impressions. The physicians of the Control group (n = 141) just reviewed the photographs once again. The P-values of top accuracies between before and after assistance of the trainees are annotated.

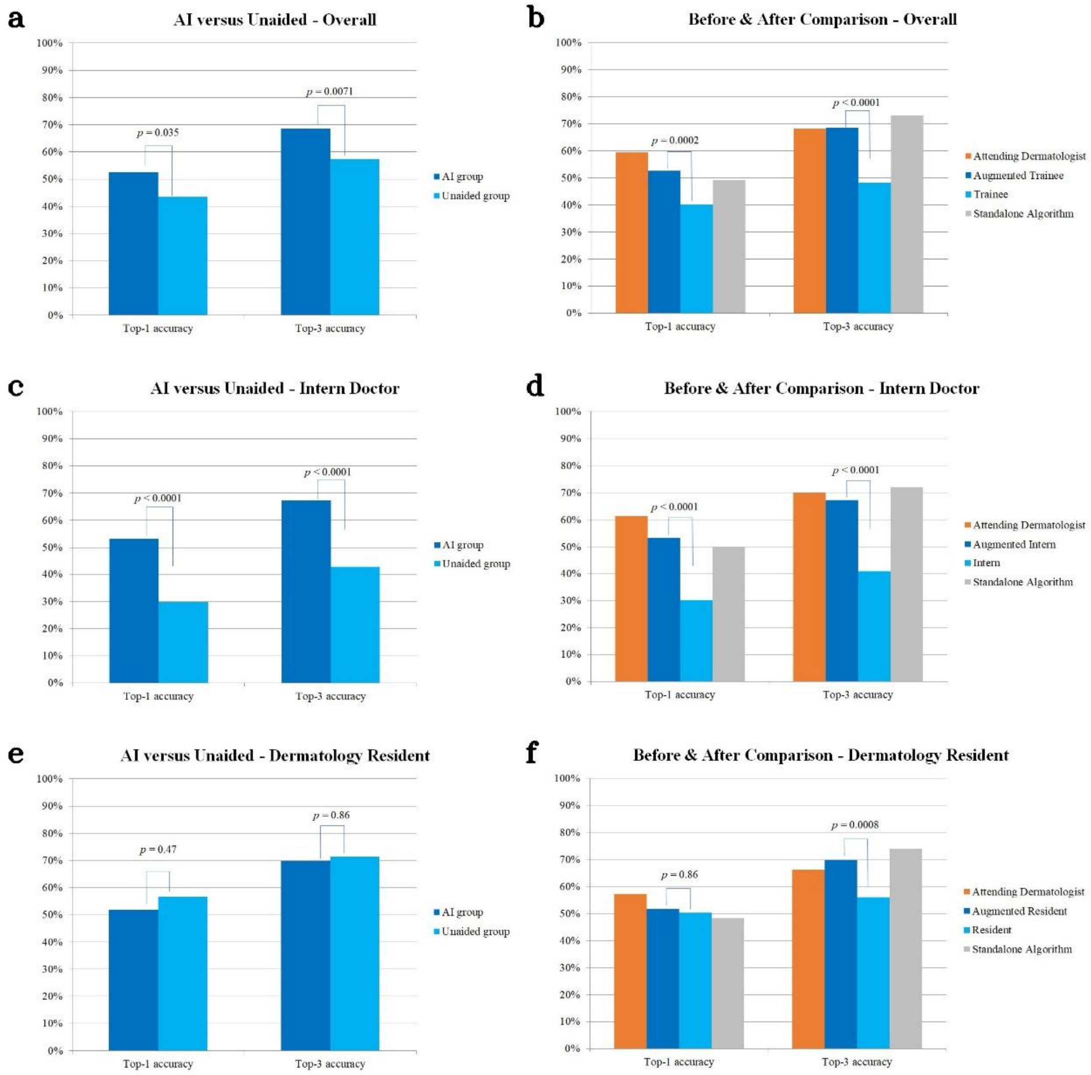


Figure 5. Top accuracies for diagnosing exact diseases in the randomized controlled trial **(A)** All Trainees – AI (N=295) versus Unaided (N=281) **(B)** All Trainees – Before & After the assistance of the algorithm in the AI group (N=295) **(C)** Interns – AI (N=150) versus Unaided (N=138) **(D)** Interns – Before & After the assistance of the algorithm in the AI group (N=150) **(E)** Dermatology Residents – AI (N=145) versus Unaided (N=143) **(F)** Dermatology Residents – Before & After the assistance of the algorithm in the AI group (N=145)

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." The P-values of Top accuracies between the AI Group and the Unaided and between before and after the assistance of the trainees are described in Supplementary table 5 and 6, respectively.

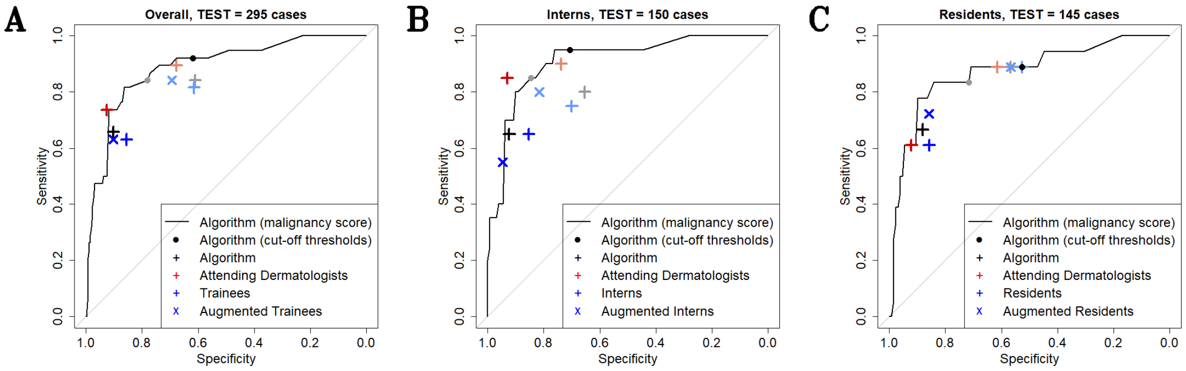


Figure 6. Sensitivity and specificity on the ROC curve for determining malignancy in the AI group in the randomized controlled trial

(A) All trainees in the AI group (N=295) **(B)** Interns in the AI group (N=150) **(C)** Dermatology Residents in the AI group (N=145)

Dark blue cross (+) – Trainees; malignancy decision derived from Top-3 predictions

Pale blue cross (+) – Trainees; malignancy decision derived from Top-1 predictions

Dark blue x-cross (×) – Augmented trainees; malignancy decision derived from Top-3 predictions

Pale blue x-cross (×) – Augmented trainees; malignancy decision derived from Top-1 predictions

Dark red cross (+) – Attending dermatologists; malignancy decision derived from Top-3 predictions

Pale red cross (+) – Attending dermatologists; malignancy decision derived from Top-1 predictions

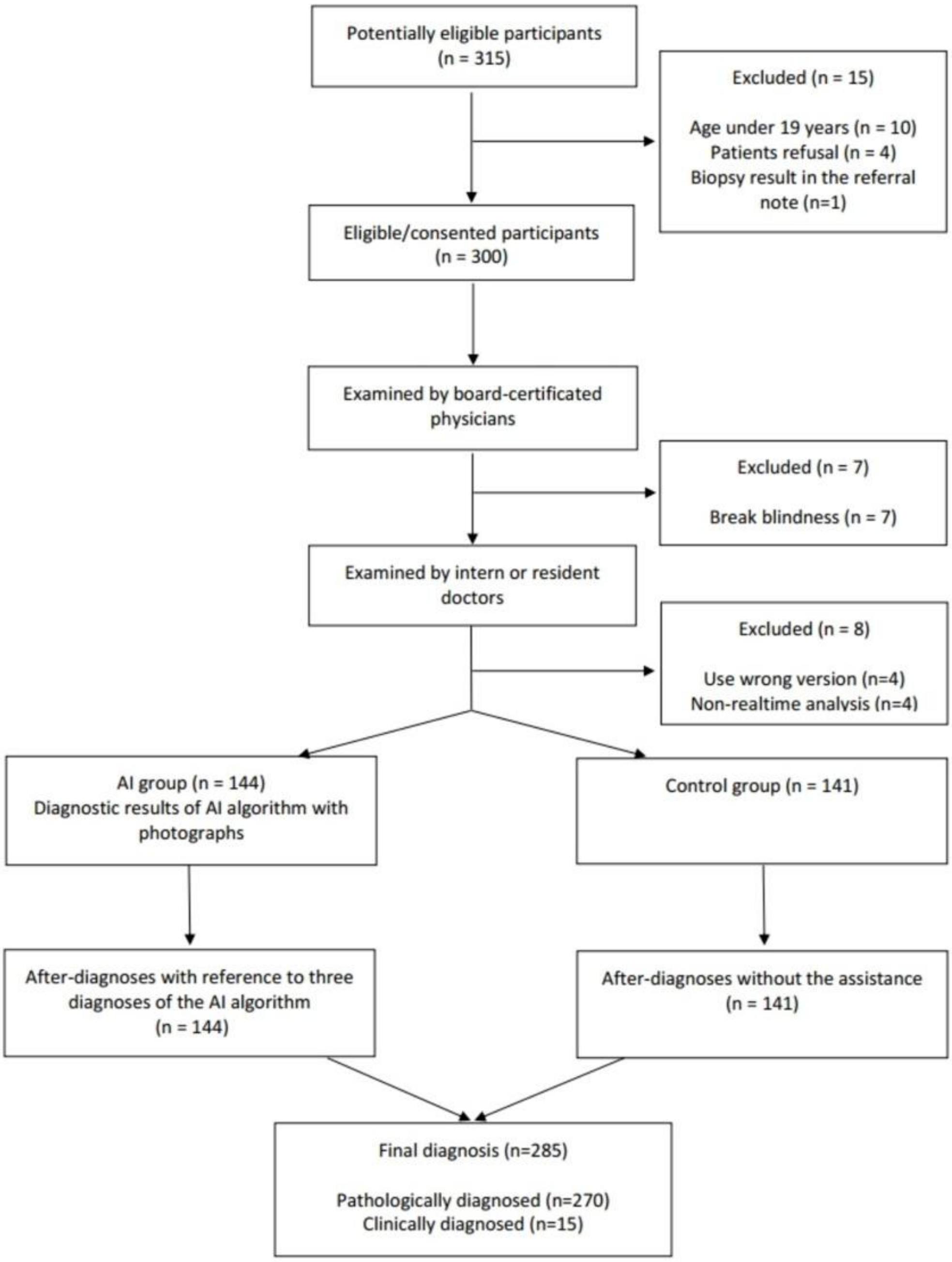
Black cross (+) – Algorithm; malignancy decision derived from Top-3 predictions

Pale black cross (+) – Algorithm; malignancy decision derived from Top-1 predictions

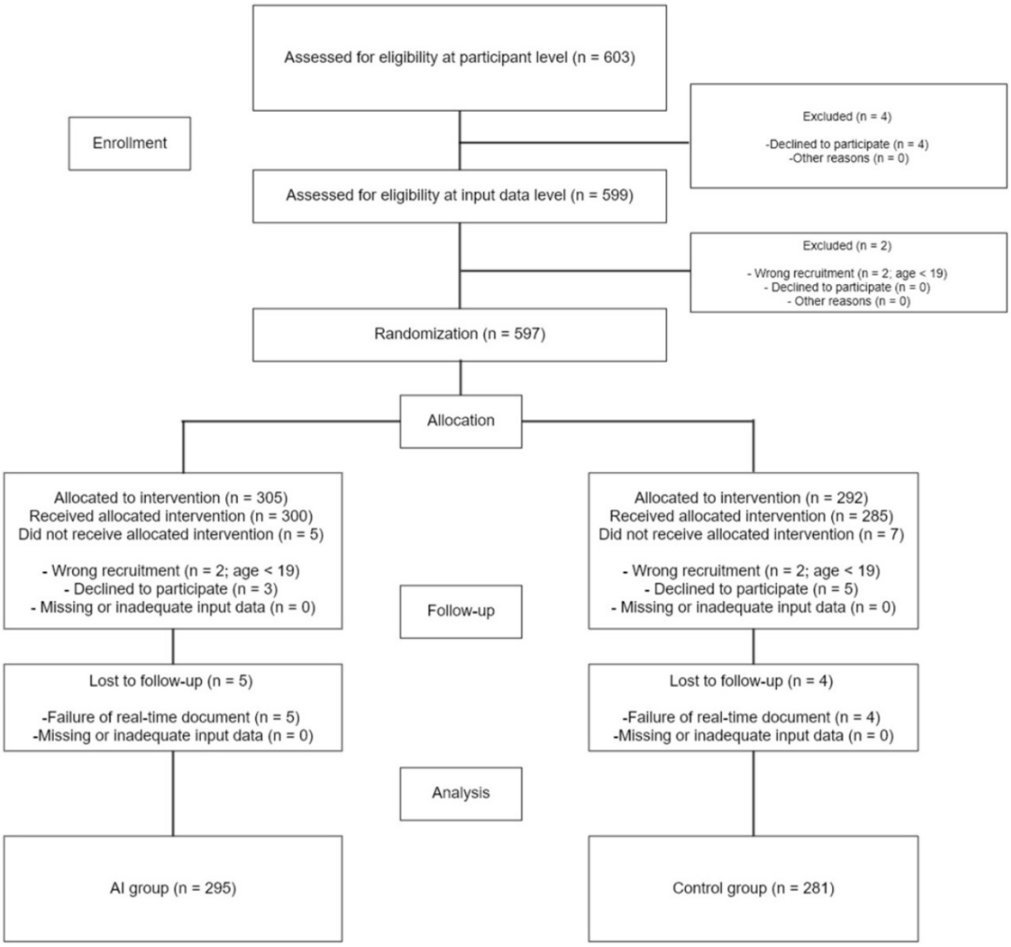
Black line – Algorithm; malignancy decision derived from the malignancy score

Black dot (●) – Algorithm at the high-sensitivity threshold

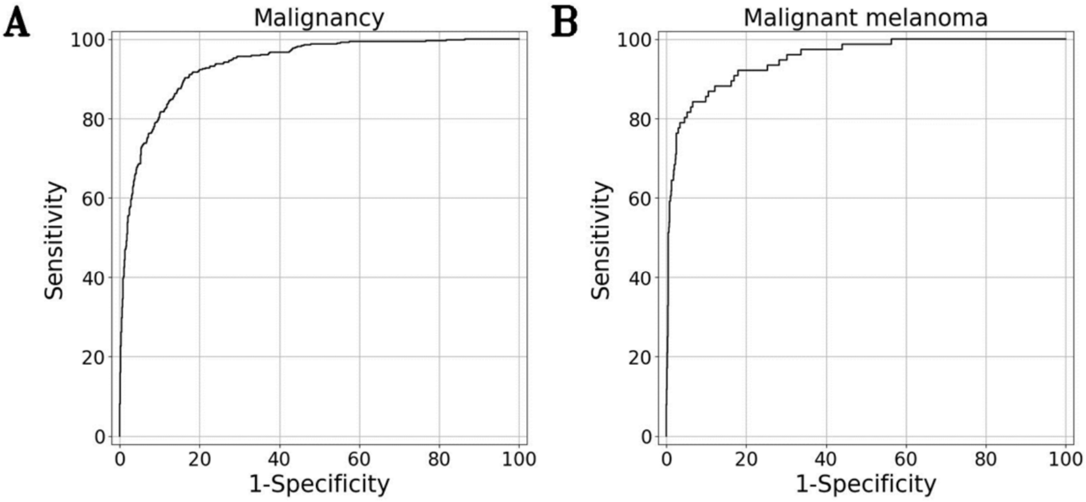
Pale black dot (●) – Algorithm at the high-specificity threshold


Potentially eligible participants (n = 315)
Excluded (n = 15)
Age under 19 years (n = 10)
Patients refusal (n = 4)
Biopsy result in the referral note (n=1)
Eligible/consented participants (n = 300)
Examined by board-certificated physicians
Excluded (n = 7)
Break blindness (n = 7)
Examined by intern or resident doctors
Excluded (n = 8)
Use wrong version (n=4)
Non-realtime analysis (n=4)
AI group (n = 144) Diagnostic results of AI algorithm with photographs
Control group (n = 141)
After-diagnoses with reference to three diagnoses of the AI algorithm (n = 144)
After-diagnoses without the assistance (n = 141)
Final diagnosis (n=285)
Pathologically diagnosed (n=270)
Clinically diagnosed (n=15)

Supplementary Figure 1. Study flowchart of the prospective controlled before-and-after study


Assessed for eligibility at participant level (n = 603)
Enrollment
Excluded (n = 4)
-Declined to participate (n = 4)
-Other reasons (n = 0)
Assessed for eligibility at input data level (n = 599)
Excluded (n = 2)
- Wrong recruitment (n = 2; age < 19)
- Declined to participate (n = 0)
- Other reasons (n = 0)
Randomization (n = 597)
Allocation
Allocated to intervention (n = 305)
Received allocated intervention (n = 300)
Did not receive allocated intervention (n = 5)
- Wrong recruitment (n = 2; age < 19)
- Declined to participate (n = 3)
- Missing or inadequate input data (n = 0)
Allocated to intervention (n = 292)
Received allocated intervention (n = 285)
Did not receive allocated intervention (n = 7)
- Wrong recruitment (n = 2; age < 19)
- Declined to participate (n = 5)
- Missing or inadequate input data (n = 0)
Follow-up
Lost to follow-up (n = 5)
-Failure of real-time document (n = 5)
-Missing or inadequate input data (n = 0)
Lost to follow-up (n = 4)
-Failure of real-time document (n = 4)
-Missing or inadequate input data (n = 0)
Analysis
AI group (n = 295)
Control group (n = 281)

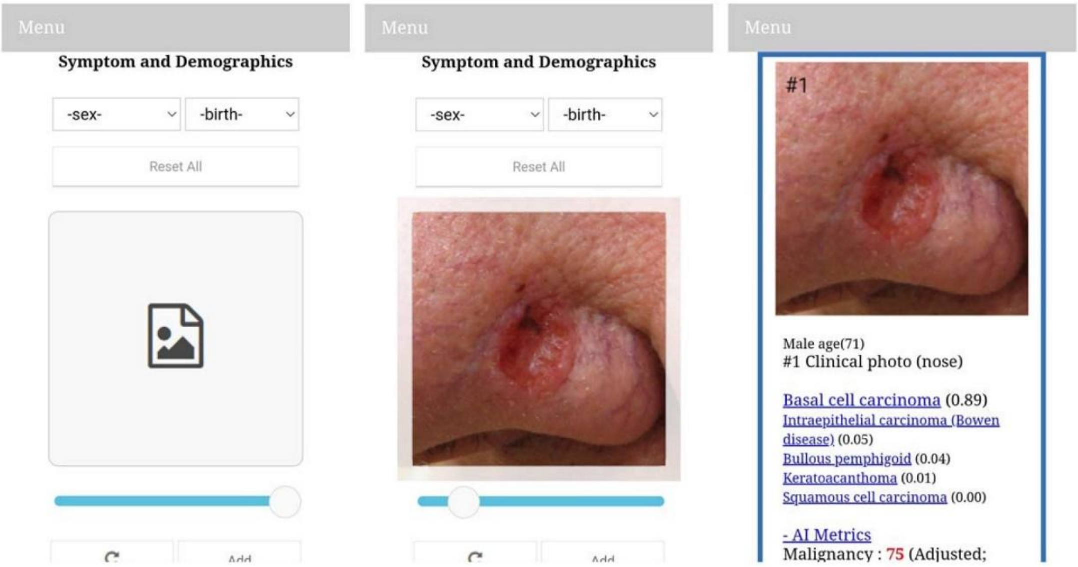
Supplementary Figure 2. Study flowchart of the randomized controlled trial



Supplementary Figure 3. 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 multi-class 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 multi-class classification.



Male age(71)
#1 Clinical photo (nose)

Basal cell carcinoma (0.89)
Intraepithelial carcinoma (Bowen disease) (0.05)
Bullous pemphigoid (0.04)
Keratoacanthoma (0.01)
Squamous cell carcinoma (0.00)

- AI Metrics
Malignancy : 75 (Adjusted;

Supplementary Figure 4. An example using the online algorithm in the randomized controlled trial

*The example photograph came from https://en.wikipedia.org/wiki/Basal-cell_carcinoma (MD, James Heilman). Algorithm's five diagnoses, their probabilities, and malignancy score were used for the experiment. The online DEMO of the algorithm is testable at https://b2020.modelderm.com/#world via PCs and mobile devices using internet browsers (Chrome and Edge browser recommended). Interpretation of the Top outputs and malignancy output was instructed as follows:

1. Top output; the Top output range from 0.0 to 1.0.

   Top output ≥ 0.2 : the predicted diagnosis is a meaningful differential diagnosis.

   Top output < 0.2 : only a small chance for the predicted diagnosis.

2. Malignancy output; the malignancy output ranges from 0 to 100.

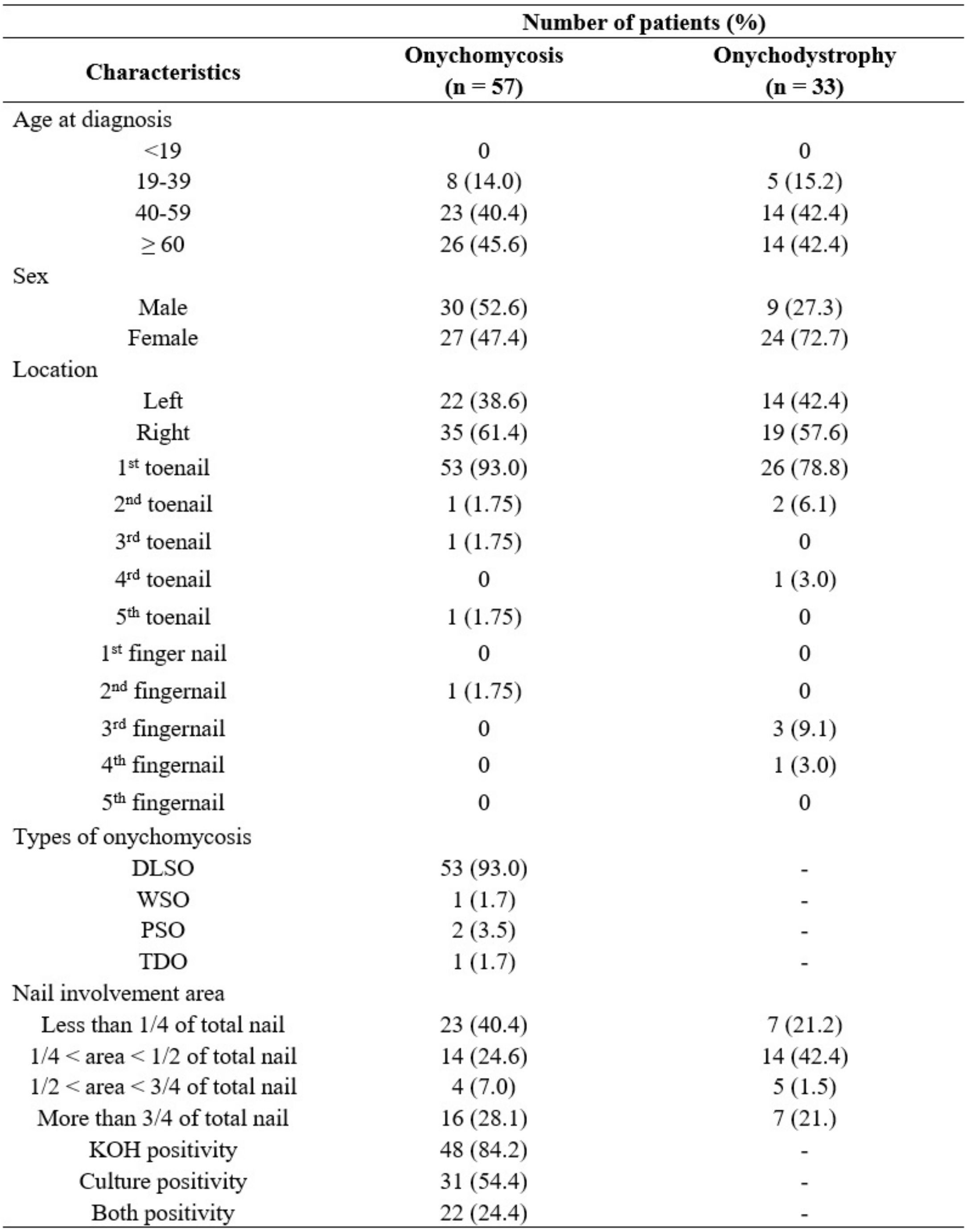
   Malignancy score ≥ 20 : High chance of malignancy

   Malignancy score ≥ 10 and < 20 : Still some chance of malignancy.

   Malignancy score < 10 : Maybe benign

TABLE LEGENDS

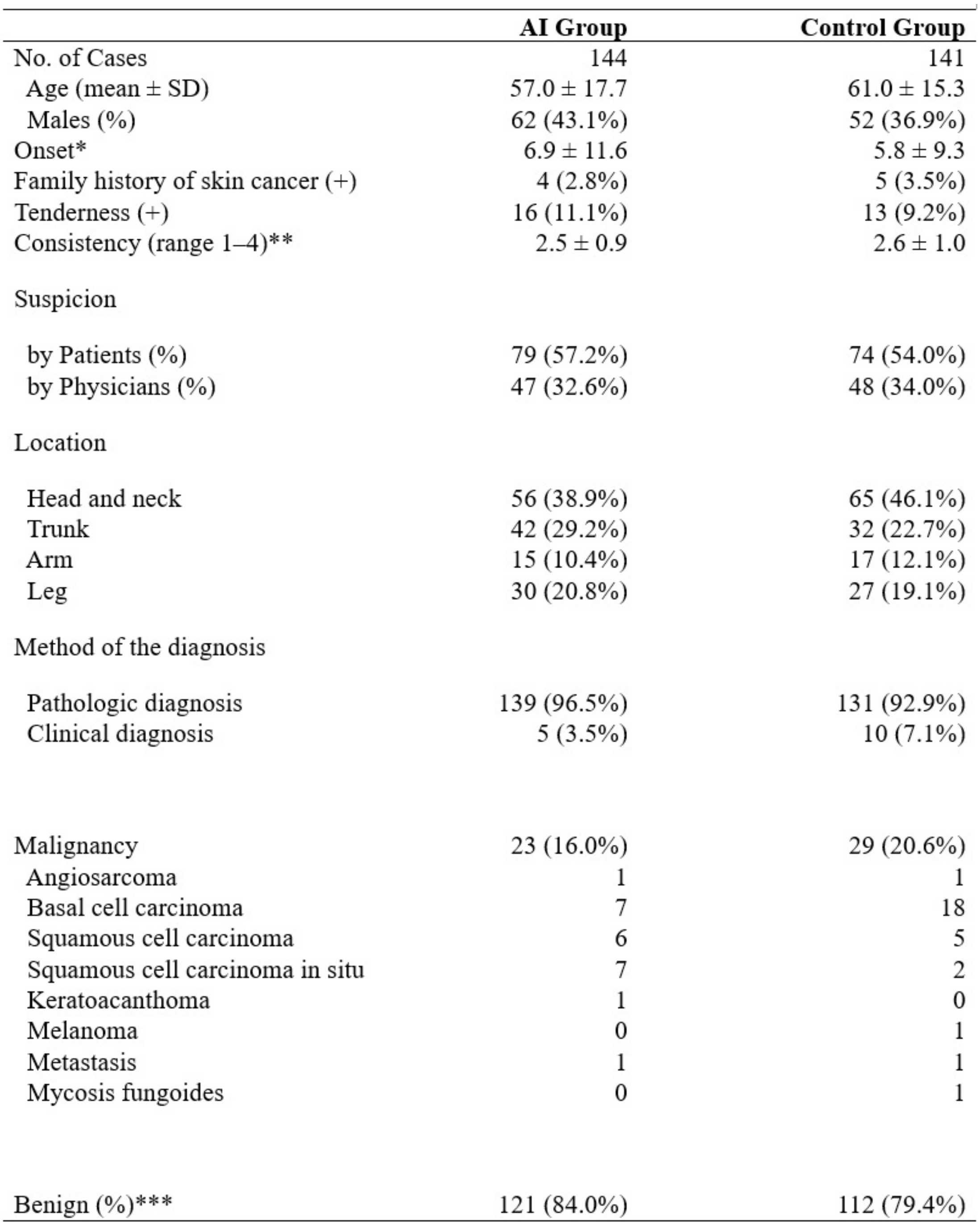
Table 1. Dataset and demographic information in the onychomycosis study



Characteristics	Number of patients (%)	
	Onychomycosis (n = 57)	**Onychodystrophy (n = 33)**
Age at diagnosis		
<19	0	0
19-39	8 (14.0)	5 (15.2)
40-59	23 (40.4)	14 (42.4)
≥ 60	26 (45.6)	14 (42.4)
Sex		
Male	30 (52.6)	9 (27.3)
Female	27 (47.4)	24 (72.7)
Location		
Left	22 (38.6)	14 (42.4)
Right	35 (61.4)	19 (57.6)
1st toenail	53 (93.0)	26 (78.8)
2nd toenail	1 (1.75)	2 (6.1)
3rd toenail	1 (1.75)	0
4rd toenail	0	1 (3.0)
5th toenail	1 (1.75)	0
1st finger nail	0	0
2nd fingernail	1 (1.75)	0
3rd fingernail	0	3 (9.1)
4th fingernail	0	1 (3.0)
5th fingernail	0	0
Types of onychomycosis		
DLSO	53 (93.0)	-
WSO	1 (1.7)	-
PSO	2 (3.5)	-
TDO	1 (1.7)	-
Nail involvement area		
Less than 1/4 of total nail	23 (40.4)	7 (21.2)
1/4 < area < 1/2 of total nail	14 (24.6)	14 (42.4)
1/2 < area < 3/4 of total nail	4 (7.0)	5 (1.5)
More than 3/4 of total nail	16 (28.1)	7 (21.)
KOH positivity	48 (84.2)	-
Culture positivity	31 (54.4)	-
Both positivity	22 (24.4)	-

Abbreviation: DLSO, distal and lateral subungual onychomycosis; WSO, white superficial onychomycosis; PSO, proximal subungual onychomycosis; TDO, total dystrophic onychomycosis

Table 2. Dataset and demographic information in the prospective before-and-after study



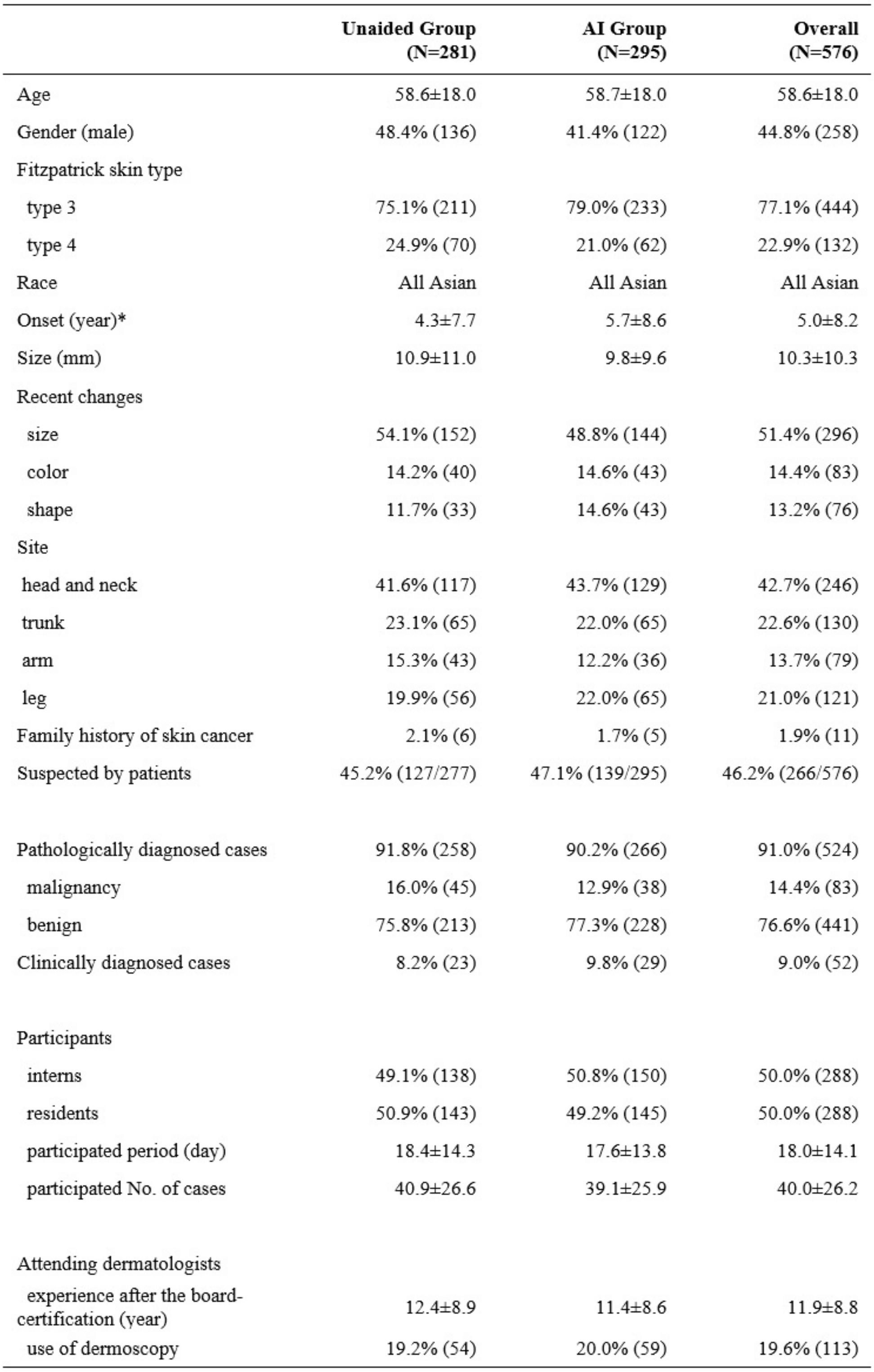
	AI Group	Control Group
No. of Cases	144	141
Age (mean ± SD)	57.0 ± 17.7	61.0 ± 15.3
Males (%)	62 (43.1%)	52 (36.9%)
Onset*	6.9 ± 11.6	5.8 ± 9.3
Family history of skin cancer (+)	4 (2.8%)	5 (3.5%)
Tenderness (+)	16 (11.1%)	13 (9.2%)
Consistency (range 1–4)**	2.5 ± 0.9	2.6 ± 1.0
Suspicion		
by Patients (%)	79 (57.2%)	74 (54.0%)
by Physicians (%)	47 (32.6%)	48 (34.0%)
Location		
Head and neck	56 (38.9%)	65 (46.1%)
Trunk	42 (29.2%)	32 (22.7%)
Arm	15 (10.4%)	17 (12.1%)
Leg	30 (20.8%)	27 (19.1%)
Method of the diagnosis		
Pathologic diagnosis	139 (96.5%)	131 (92.9%)
Clinical diagnosis	5 (3.5%)	10 (7.1%)
Malignancy	23 (16.0%)	29 (20.6%)
Angiosarcoma	1	1
Basal cell carcinoma	7	18
Squamous cell carcinoma	6	5
Squamous cell carcinoma in situ	7	2
Keratoacanthoma	1	0
Melanoma	0	1
Metastasis	1	1
Mycosis fungoides	0	1
Benign (%)***	121 (84.0%)	112 (79.4%)

* Onset were available in 93.3% of cases (266 cases).

** The consistency was annotated as follows: 1 = hard, 2 = renitent, 3 = normal, and 4 = soft.

*** The details of the benign conditions are listed in the Supplementary table 1.

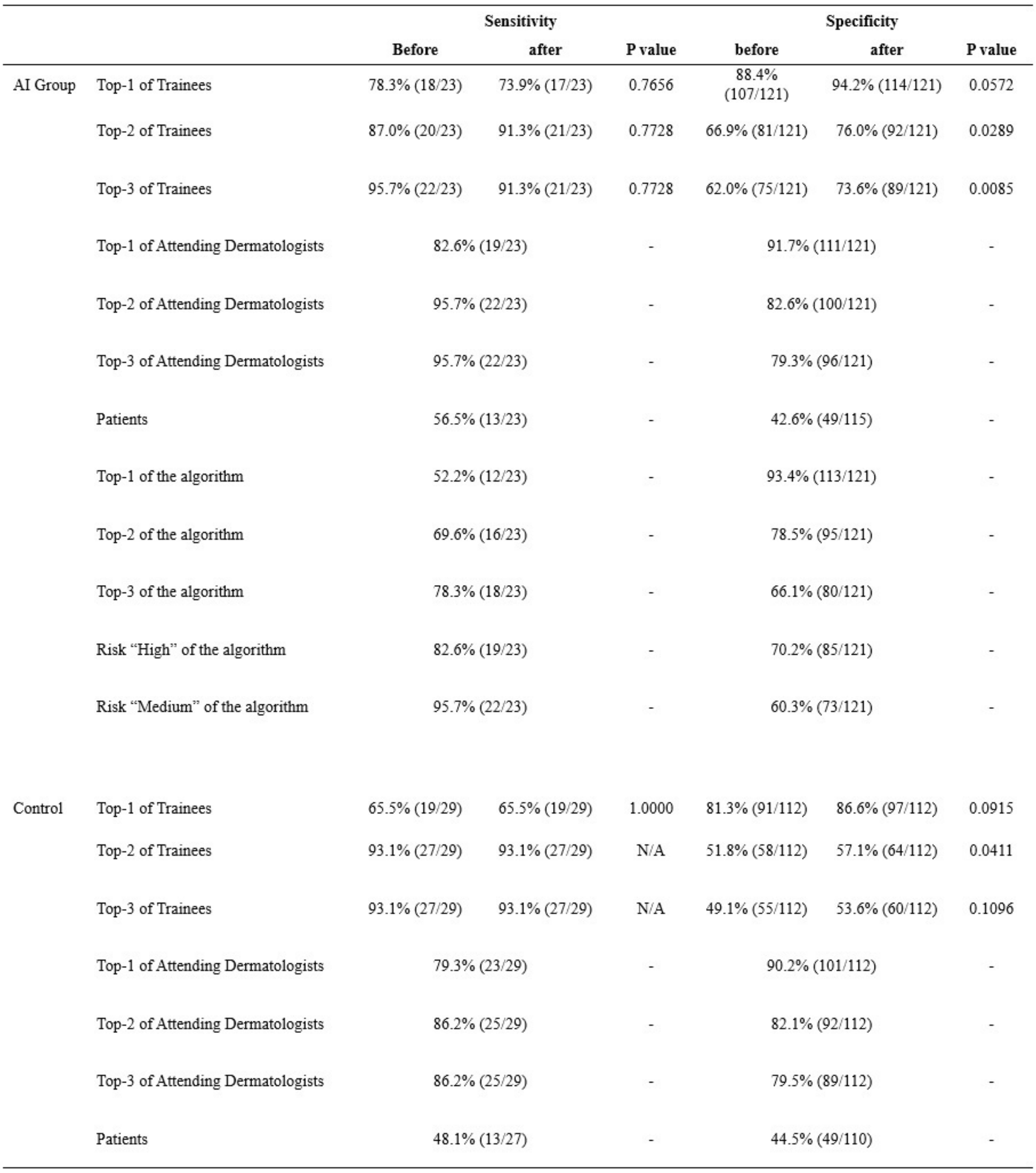
Table 3. Demographics and the status of the randomization in the randomized controlled trial



	Unaided Group (N=281)	**AI Group (N=295)**	**Overall (N=576)**
Age	58.6±18.0	58.7±18.0	58.6±18.0
Gender (male)	48.4% (136)	41.4% (122)	44.8% (258)
Fitzpatrick skin type			
type 3	75.1% (211)	79.0% (233)	77.1% (444)
type 4	24.9% (70)	21.0% (62)	22.9% (132)
Race	All Asian	All Asian	All Asian
Onset (year)*	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)	12.9% (38)	14.4% (83)
benign	75.8% (213)	77.3% (228)	76.6% (441)
Clinically diagnosed cases	8.2% (23)	9.8% (29)	9.0% (52)
Participants			
interns	49.1% (138)	50.8% (150)	50.0% (288)
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 (year)	12.4±8.9	11.4±8.6	11.9±8.8
use of dermoscopy	19.2% (54)	20.0% (59)	19.6% (113)

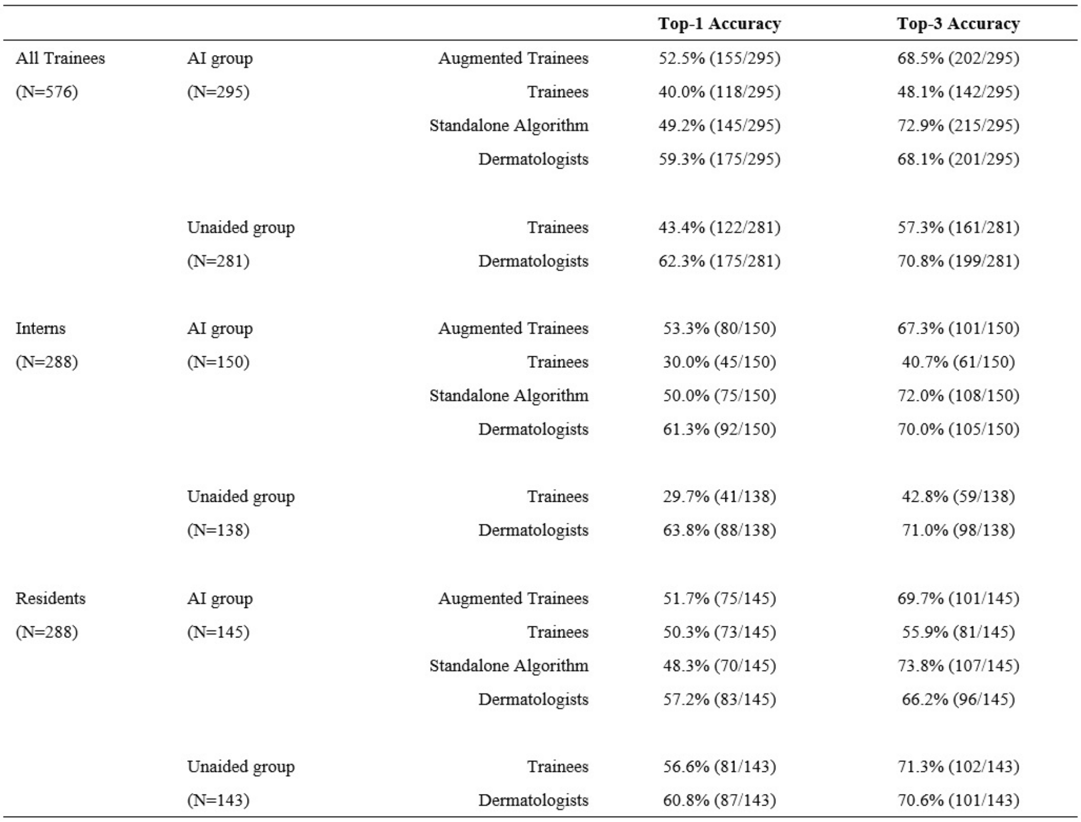
* A total of 88.6% (249 cases), 90.2% (266 cases), and 89.4% (515 cases) onset records were available in the Unaided group, AI group, and the Overall, respectively.

Table 4. Summaries of the sensitivity and specificity in the prospective before-and-after study



		Sensitivity			Specificity		
		Before	after	P value	before	after	P value
AI Group	Top-1 of Trainees	78.3% (18/23)	73.9% (17/23)	0.7656	88.4% (107/121)	94.2% (114/121)	0.0572
	Top-2 of Trainees	87.0% (20/23)	91.3% (21/23)	0.7728	66.9% (81/121)	76.0% (92/121)	0.0289
	Top-3 of Trainees	95.7% (22/23)	91.3% (21/23)	0.7728	62.0% (75/121)	73.6% (89/121)	0.0085
	Top-1 of Attending Dermatologists	82.6% (19/23)		-	91.7% (111/121)		-
	Top-2 of Attending Dermatologists	95.7% (22/23)		-	82.6% (100/121)		-
	Top-3 of Attending Dermatologists	95.7% (22/23)		-	79.3% (96/121)		-
	Patients	56.5% (13/23)		-	42.6% (49/115)		-
	Top-1 of the algorithm	52.2% (12/23)		-	93.4% (113/121)		-
	Top-2 of the algorithm	69.6% (16/23)		-	78.5% (95/121)		-
	Top-3 of the algorithm	78.3% (18/23)		-	66.1% (80/121)		-
	Risk "High" of the algorithm	82.6% (19/23)		-	70.2% (85/121)		-
	Risk "Medium" of the algorithm	95.7% (22/23)		-	60.3% (73/121)		-
Control	Top-1 of Trainees	65.5% (19/29)	65.5% (19/29)	1.0000	81.3% (91/112)	86.6% (97/112)	0.0915
	Top-2 of Trainees	93.1% (27/29)	93.1% (27/29)	N/A	51.8% (58/112)	57.1% (64/112)	0.0411
	Top-3 of Trainees	93.1% (27/29)	93.1% (27/29)	N/A	49.1% (55/112)	53.6% (60/112)	0.1096
	Top-1 of Attending Dermatologists	79.3% (23/29)		-	90.2% (101/112)		-
	Top-2 of Attending Dermatologists	86.2% (25/29)		-	82.1% (92/112)		-
	Top-3 of Attending Dermatologists	86.2% (25/29)		-	79.5% (89/112)		-
	Patients	48.1% (13/27)		-	44.5% (49/110)		-

Table 5. Top-1 and Top-3 accuracy of the participants and algorithm in the randomized controlled trial

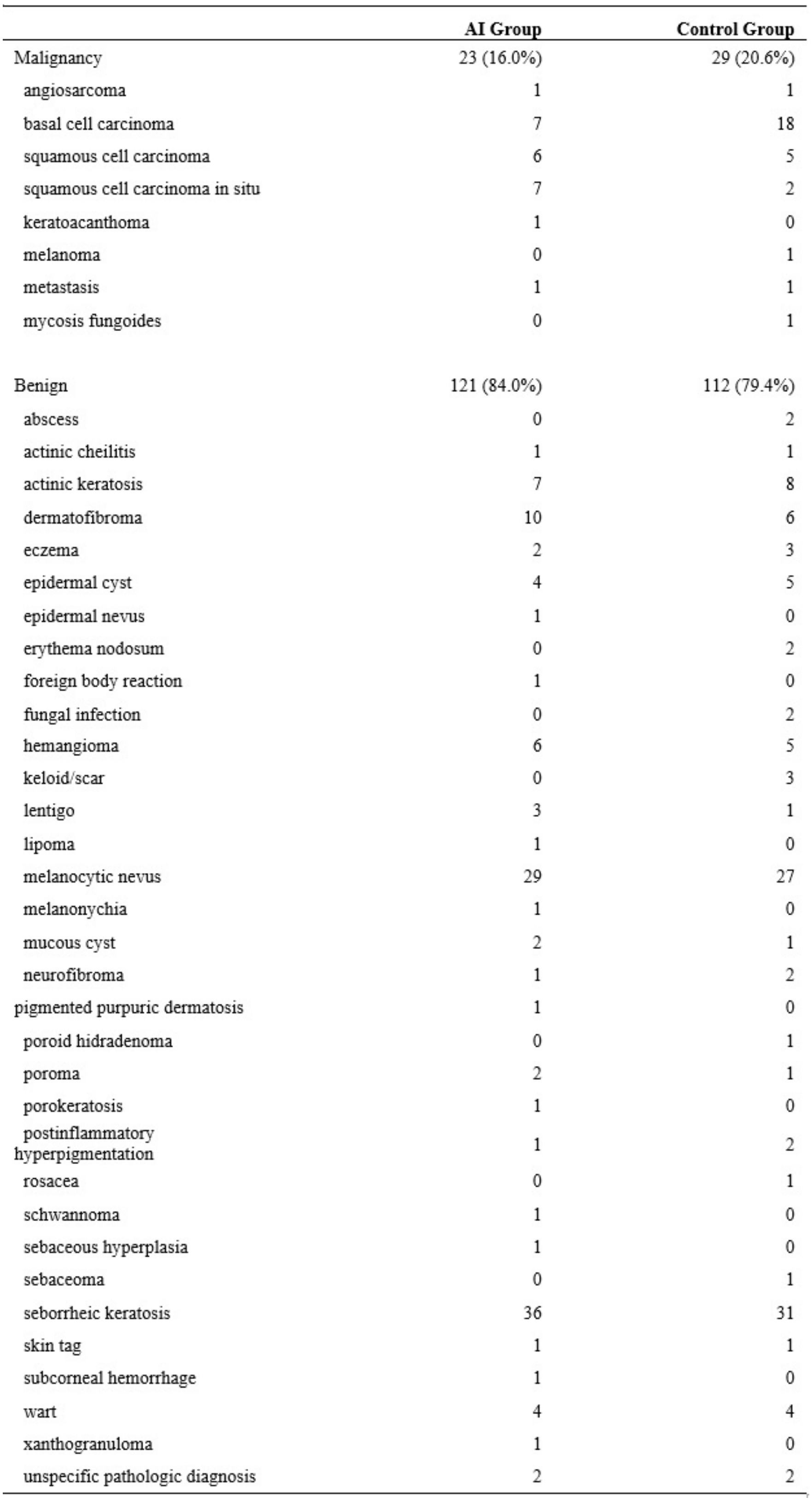


			Top-1 Accuracy	Top-3 Accuracy
All Trainees	AI group	Augmented Trainees	52.5% (155/295)	68.5% (202/295)
(N=576)	(N=295)	Trainees	40.0% (118/295)	48.1% (142/295)
		Standalone Algorithm	49.2% (145/295)	72.9% (215/295)
		Dermatologists	59.3% (175/295)	68.1% (201/295)
	Unaided group	Trainees	43.4% (122/281)	57.3% (161/281)
	(N=281)	Dermatologists	62.3% (175/281)	70.8% (199/281)
Interns	AI group	Augmented Trainees	53.3% (80/150)	67.3% (101/150)
(N=288)	(N=150)	Trainees	30.0% (45/150)	40.7% (61/150)
		Standalone Algorithm	50.0% (75/150)	72.0% (108/150)
		Dermatologists	61.3% (92/150)	70.0% (105/150)
	Unaided group	Trainees	29.7% (41/138)	42.8% (59/138)
	(N=138)	Dermatologists	63.8% (88/138)	71.0% (98/138)
Residents	AI group	Augmented Trainees	51.7% (75/145)	69.7% (101/145)
(N=288)	(N=145)	Trainees	50.3% (73/145)	55.9% (81/145)
		Standalone Algorithm	48.3% (70/145)	73.8% (107/145)
		Dermatologists	57.2% (83/145)	66.2% (96/145)
	Unaided group	Trainees	56.6% (81/143)	71.3% (102/143)
	(N=143)	Dermatologists	60.8% (87/143)	70.6% (101/143)

Table 6. Sensitivity, specificity, positive predictive value, and negative predictive value derived from Top-1 and Top-3 predictions in the randomized controlled trial

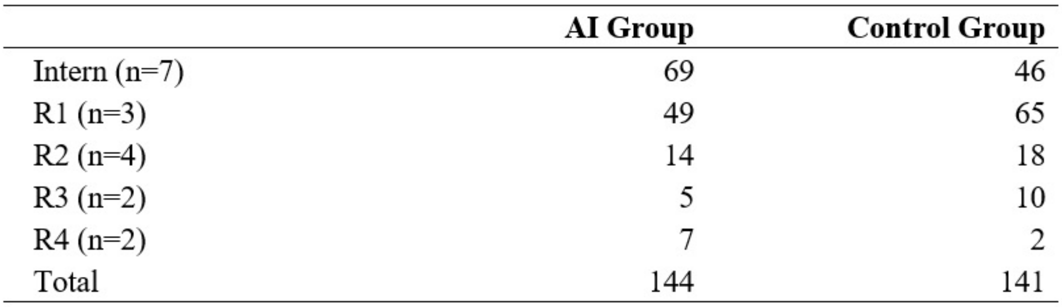
			Sensitivity / Specificity from Top-1 prediction	**Sensitivity / Specificity from Top-3 predictions**	**PPV / NPV from Top-1 prediction**	**PPV / NPV from Top-3 predictions**
All Trainees	AI group	Augmented Trainees	63.2% (24/38) / 90.3% (232/257)	84.2% (32/38) / 69.3% (178/257)	49.0% (24/49) / 94.3% (232/246)	28.8% (32/111) / 96.7% (178/184)
(N=576)	(N=295)	Trainees	63.2% (24/38) / 85.6% (220/257)	81.6% (31/38) / 61.5% (158/257)	39.3% (24/61) / 94.0% (220/234)	23.8% (31/130) / 95.8% (158/165)
		Standalone Algorithm	65.8% (25/38) / 90.3% (232/257)	84.2% (32/38) / 61.1% (157/257)	50.0% (25/50) / 94.7% (232/245)	24.2% (32/132) / 96.3% (157/163)
		Dermatologists	73.7% (28/38) / 92.6% (238/257)	89.5% (34/38) / 67.7% (174/257)	59.6% (28/47) / 96.0% (238/248)	29.1% (34/117) / 97.8% (174/178)
	Unaided group	Trainees	53.3% (24/45) / 85.2% (201/236)	75.6% (34/45) / 63.1% (149/236)	40.7% (24/59) / 90.5% (201/222)	28.1% (34/121) / 93.1% (149/160)
	(N=281)	Dermatologists	64.4% (29/45) / 91.5% (216/236)	88.9% (40/45) / 67.4% (159/236)	59.2% (29/49) / 93.1% (216/232)	34.2% (40/117) / 97.0% (159/164)
Interns	AI group	Augmented Trainees	55.0% (11/20) / 94.6% (123/130)	80.0% (16/20) / 81.5% (106/130)	61.1% (11/18) / 93.2% (123/132)	40.0% (16/40) / 96.4% (106/110)
(N=288)	(N=150)	Trainees	65.0% (13/20) / 85.4% (111/130)	75.0% (15/20) / 70.0% (91/130)	40.6% (13/32) / 94.1% (111/118)	27.8% (15/54) / 94.8% (91/96)
		Standalone Algorithm	65.0% (13/20) / 92.3% (120/130)	80.0% (16/20) / 65.4% (85/130)	56.5% (13/23) / 94.5% (120/127)	26.2% (16/61) / 95.5% (85/89)
		Dermatologists	85.0% (17/20) / 93.1% (121/130)	90.0% (18/20) / 73.8% (96/130)	65.4% (17/26) / 97.6% (121/124)	34.6% (18/52) / 98.0% (96/98)
	Unaided group	Trainees	43.8% (7/16) / 83.6% (102/122)	56.3% (9/16) / 68.9% (84/122)	25.9% (7/27) / 91.9% (102/111)	19.1% (9/47) / 92.3% (84/91)
	(N=138)	Dermatologists	50.0% (8/16) / 95.1% (116/122)	93.8% (15/16) / 68.0% (83/122)	57.1% (8/14) / 93.5% (116/124)	27.8% (15/54) / 98.8% (83/84)
Residents	AI group	Augmented Trainees	72.2% (13/18) / 85.8% (109/127)	88.9% (16/18) / 56.7% (72/127)	41.9% (13/31) / 95.6% (109/114)	22.5% (16/71) / 97.3% (72/74)
(N=288)	(N=145)	Trainees	61.1% (11/18) / 85.8% (109/127)	88.9% (16/18) / 52.8% (67/127)	37.9% (11/29) / 94.0% (109/116)	21.1% (16/76) / 97.1% (67/69)
		Standalone Algorithm	66.7% (12/18) / 88.2% (112/127)	88.9% (16/18) / 56.7% (72/127)	44.4% (12/27) / 94.9% (112/118)	22.5% (16/71) / 97.3% (72/74)
		Dermatologists	61.1% (11/18) / 92.1% (117/127)	88.9% (16/18) / 61.4% (78/127)	52.4% (11/21) / 94.4% (117/124)	24.6% (16/65) / 97.5% (78/80)
	Unaided group	Trainees	58.6% (17/29) / 86.8% (99/114)	86.2% (25/29) / 57.0% (65/114)	53.1% (17/32) / 89.2% (99/111)	33.8% (25/74) / 94.2% (65/69)
	(N=143)	Dermatologists	72.4% (21/29) / 87.7% (100/114)	86.2% (25/29) / 66.7% (76/114)	60.0% (21/35) / 92.6% (100/108)	39.7% (25/63) / 95.0% (76/80)

Supplementary Table 1. Benign and malignant disease dataset in the prospective before-and-after study



	AI Group	Control Group
Malignancy	23 (16.0%)	29 (20.6%)
angiosarcoma	1	1
basal cell carcinoma	7	18
squamous cell carcinoma	6	5
squamous cell carcinoma in situ	7	2
keratoacanthoma	1	0
melanoma	0	1
metastasis	1	1
mycosis fungoides	0	1
Benign	121 (84.0%)	112 (79.4%)
abscess	0	2
actinic cheilitis	1	1
actinic keratosis	7	8
dermatofibroma	10	6
eczema	2	3
epidermal cyst	4	5
epidermal nevus	1	0
erythema nodosum	0	2
foreign body reaction	1	0
fungal infection	0	2
hemangioma	6	5
keloid/scar	0	3
lentigo	3	1
lipoma	1	0
melanocytic nevus	29	27
melanonychia	1	0
mucous cyst	2	1
neurofibroma	1	2
pigmented purpuric dermatosis	1	0
poroid hidradenoma	0	1
poroma	2	1
porokeratosis	1	0
postinflammatory hyperpigmentation	1	2
rosacea	0	1
schwannoma	1	0
sebaceous hyperplasia	1	0
sebaceoma	0	1
seborrheic keratosis	36	31
skin tag	1	1
subcorneal hemorrhage	1	0
wart	4	4
xanthogranuloma	1	0
unspecific pathologic diagnosis	2	2

Supplementary Table 2. Number of examined cases and the grade of the participants in the prospective before-and-after study



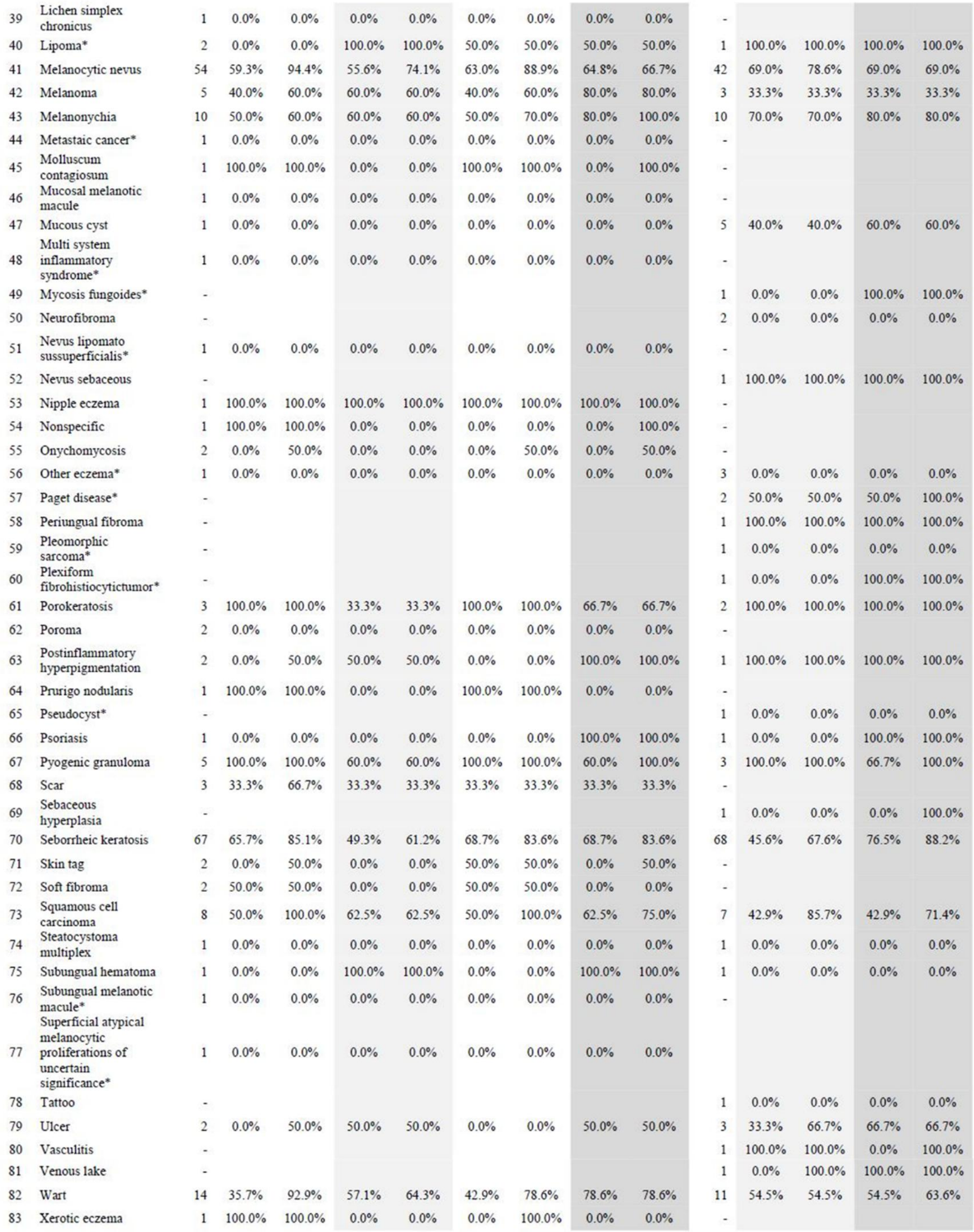
	AI Group	**Control Group**
Intern (n=7)	69	46
R1 (n=3)	49	65
R2 (n=4)	14	18
R3 (n=2)	5	10
R4 (n=2)	7	2
Total	144	141

The R1 represents dermatology the first-year resident trainee.

Supplementary Table 3. Top-1 and Top-3 accuracy for the 83 conditions in the randomized controlled trial

			AI Group									Unaided Group			
			Stanalone Algorithm		Trainee		Augmented Trainee		Attending DER			Trainee		Attending DER	
		No of cases	Top1	Top3	Top1	Top3	Top1	Top3	Top1	Top3	No of cases	Top1	Top3	Top1	Top3
1	Actinic keratosis	14	64.3%	85.7%	35.7%	50.0%	64.3%	78.6%	57.1%	78.6%	10	30.0%	60.0%	40.0%	60.0%
2	Adenoid cystic carcinoma*	-									1	0.0%	0.0%	0.0%	0.0%
3	Angiokeratoma	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
4	Angiolipoma*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	-				
5	Atypical meningioma*	-									1	0.0%	0.0%	0.0%	0.0%
6	Bartholin cyst*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
7	Basal cell carcinoma	15	53.3%	66.7%	33.3%	40.0%	60.0%	73.3%	46.7%	66.7%	19	42.1%	68.4%	78.9%	89.5%
8	Blue nevus	-									2	100.0%	100.0%	100.0%	100.0%
9	Bowen disease	7	14.3%	28.6%	14.3%	14.3%	14.3%	28.6%	28.6%	28.6%	11	9.1%	18.2%	27.3%	45.5%
10	Calcinosis cutis*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	-				
11	Callus	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	-				
12	Cheilitis	4	50.0%	50.0%	25.0%	25.0%	50.0%	50.0%	75.0%	75.0%	3	100.0%	100.0%	66.7%	66.7%
13	Chronic eczema	1	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
14	Congenital nevus	3	33.3%	33.3%	0.0%	0.0%	33.3%	33.3%	100.0%	100.0%	3	66.7%	66.7%	100.0%	100.0%
15	Cutaneous T cell lymphoma*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
16	Cyst	-									1	0.0%	0.0%	100.0%	100.0%
17	Deep fungal infection*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
18	Dermatofibroma	7	85.7%	100.0%	0.0%	0.0%	85.7%	85.7%	85.7%	85.7%	9	22.2%	44.4%	77.8%	88.9%
19	Drug eruption	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	1	0.0%	0.0%	0.0%	0.0%
20	Dysplastic nevus	5	0.0%	80.0%	20.0%	20.0%	0.0%	60.0%	100.0%	100.0%	2	0.0%	100.0%	50.0%	50.0%
21	Enchondroma*	-									1	0.0%	0.0%	0.0%	0.0%
22	Epidermal cyst	11	63.6%	100.0%	54.5%	63.6%	90.9%	90.9%	63.6%	63.6%	16	50.0%	62.5%	75.0%	75.0%
23	Fibrokeratoma*	-									1	0.0%	0.0%	0.0%	0.0%
24	Fibroma*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	-				
25	Fibrotic pseudocyst*	-									1	0.0%	0.0%	0.0%	0.0%
26	Folliculitis	-									1	0.0%	0.0%	100.0%	100.0%
27	Foreign body granuloma*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
28	Foreign body reaction*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
29	Ganglion cyst*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
30	Glomus tumor*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	-				
31	Gout*	-									1	0.0%	0.0%	0.0%	0.0%
32	Hemangioma	8	12.5%	37.5%	12.5%	25.0%	25.0%	37.5%	25.0%	25.0%	7	28.6%	28.6%	42.9%	57.1%
33	Hematoma	1	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	1	0.0%	0.0%	100.0%	100.0%
34	Kaposi sarcoma*	-									1	0.0%	0.0%	100.0%	100.0%
35	Keratoacanthoma	1	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	1	0.0%	0.0%	100.0%	100.0%
36	Leiomyosarcoma*	-									1	0.0%	0.0%	0.0%	0.0%
37	Lentigo	4	50.0%	50.0%	25.0%	25.0%	50.0%	50.0%	25.0%	25.0%	8	0.0%	0.0%	0.0%	12.5%
38	Lichen planus	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	2	0.0%	0.0%	50.0%	50.0%

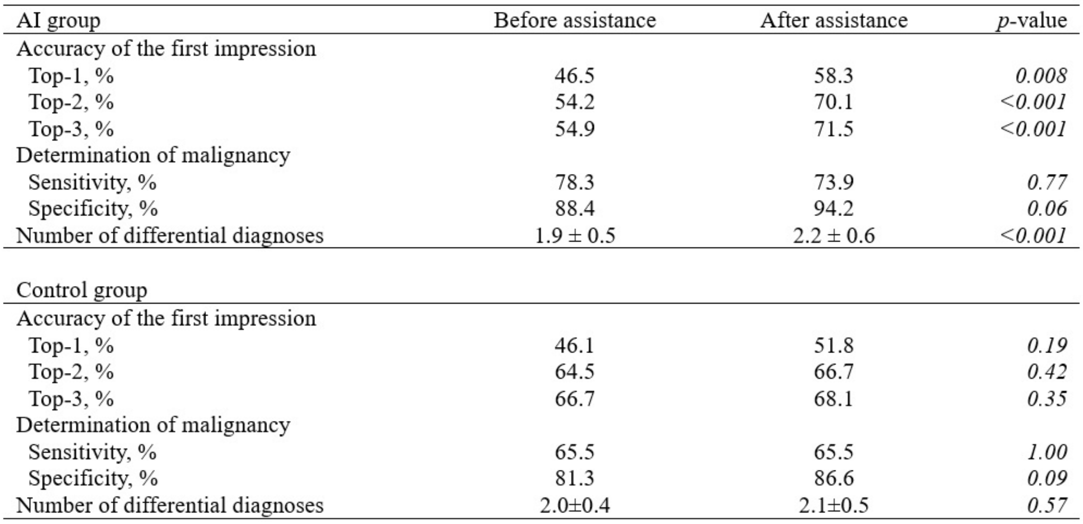
Supplementary Table 3 (continued). Top-1 and Top-3 accuracy for the 83 conditions in the randomized controlled trial



39	Lichen simplex chronicus	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
40	Lipoma*	2	0.0%	0.0%	100.0%	100.0%	50.0%	50.0%	50.0%	50.0%	1	100.0%	100.0%	100.0%	100.0%
41	Melanocytic nevus	54	59.3%	94.4%	55.6%	74.1%	63.0%	88.9%	64.8%	66.7%	42	69.0%	78.6%	69.0%	69.0%
42	Melanoma	5	40.0%	60.0%	60.0%	60.0%	40.0%	60.0%	80.0%	80.0%	3	33.3%	33.3%	33.3%	33.3%
43	Melanonychia	10	50.0%	60.0%	60.0%	60.0%	50.0%	70.0%	80.0%	100.0%	10	70.0%	70.0%	80.0%	80.0%
44	Metastaic cancer*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
45	Molluscum contagiosum	1	100.0%	100.0%	0.0%	0.0%	100.0%	100.0%	0.0%	100.0%	-				
46	Mucosal melanotic macule	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
47	Mucous cyst	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5	40.0%	40.0%	60.0%	60.0%
48	Multi system inflammatory syndrome*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
49	Mycosis fungoides*	-									1	0.0%	0.0%	100.0%	100.0%
50	Neurofibroma	-									2	0.0%	0.0%	0.0%	0.0%
51	Nevus lipomato sussuperficialis*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
52	Nevus sebaceous	-									1	100.0%	100.0%	100.0%	100.0%
53	Nipple eczema	1	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	-				
54	Nonspecific	1	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	-				
55	Onychomycosis	2	0.0%	50.0%	0.0%	0.0%	0.0%	50.0%	0.0%	50.0%	-				
56	Other eczema*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3	0.0%	0.0%	0.0%	0.0%
57	Paget disease*	-									2	50.0%	50.0%	50.0%	100.0%
58	Periungual fibroma	-									1	100.0%	100.0%	100.0%	100.0%
59	Pleomorphic sarcoma*	-									1	0.0%	0.0%	0.0%	0.0%
60	Plexiform fibrohistiocytictumor*	-									1	0.0%	0.0%	100.0%	100.0%
61	Porokeratosis	3	100.0%	100.0%	33.3%	33.3%	100.0%	100.0%	66.7%	66.7%	2	100.0%	100.0%	100.0%	100.0%
62	Poroma	2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
63	Postinflammatory hyperpigmentation	2	0.0%	50.0%	50.0%	50.0%	0.0%	0.0%	100.0%	100.0%	1	100.0%	100.0%	100.0%	100.0%
64	Prurigo nodularis	1	100.0%	100.0%	0.0%	0.0%	100.0%	100.0%	0.0%	0.0%	-				
65	Pseudocyst*	-									1	0.0%	0.0%	0.0%	0.0%
66	Psoriasis	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%	1	0.0%	0.0%	100.0%	100.0%
67	Pyogenic granuloma	5	100.0%	100.0%	60.0%	60.0%	100.0%	100.0%	60.0%	100.0%	3	100.0%	100.0%	66.7%	100.0%
68	Scar	3	33.3%	66.7%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	-				
69	Sebaceous hyperplasia	-									1	0.0%	0.0%	0.0%	100.0%
70	Seborrheic keratosis	67	65.7%	85.1%	49.3%	61.2%	68.7%	83.6%	68.7%	83.6%	68	45.6%	67.6%	76.5%	88.2%
71	Skin tag	2	0.0%	50.0%	0.0%	0.0%	50.0%	50.0%	0.0%	50.0%	-				
72	Soft fibroma	2	50.0%	50.0%	0.0%	0.0%	50.0%	50.0%	0.0%	0.0%	-				
73	Squamous cell carcinoma	8	50.0%	100.0%	62.5%	62.5%	50.0%	100.0%	62.5%	75.0%	7	42.9%	85.7%	42.9%	71.4%
74	Steatocystoma multiplex	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1	0.0%	0.0%	0.0%	0.0%
75	Subungual hematoma	1	0.0%	0.0%	100.0%	100.0%	0.0%	0.0%	100.0%	100.0%	1	0.0%	0.0%	0.0%	0.0%
76	Subungual melanotic macule*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-				
77	Superficial atypical melanocytic proliferations of uncertain significance*	1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%					
78	Tattoo	-									1	0.0%	0.0%	0.0%	0.0%
79	Ulcer	2	0.0%	50.0%	50.0%	50.0%	0.0%	0.0%	50.0%	50.0%	3	33.3%	66.7%	66.7%	66.7%
80	Vasculitis	-									1	100.0%	100.0%	0.0%	100.0%
81	Venous lake	-									1	0.0%	100.0%	100.0%	100.0%
82	Wart	14	35.7%	92.9%	57.1%	64.3%	42.9%	78.6%	78.6%	78.6%	11	54.5%	54.5%	54.5%	63.6%
83	Xerotic eczema	1	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	-				

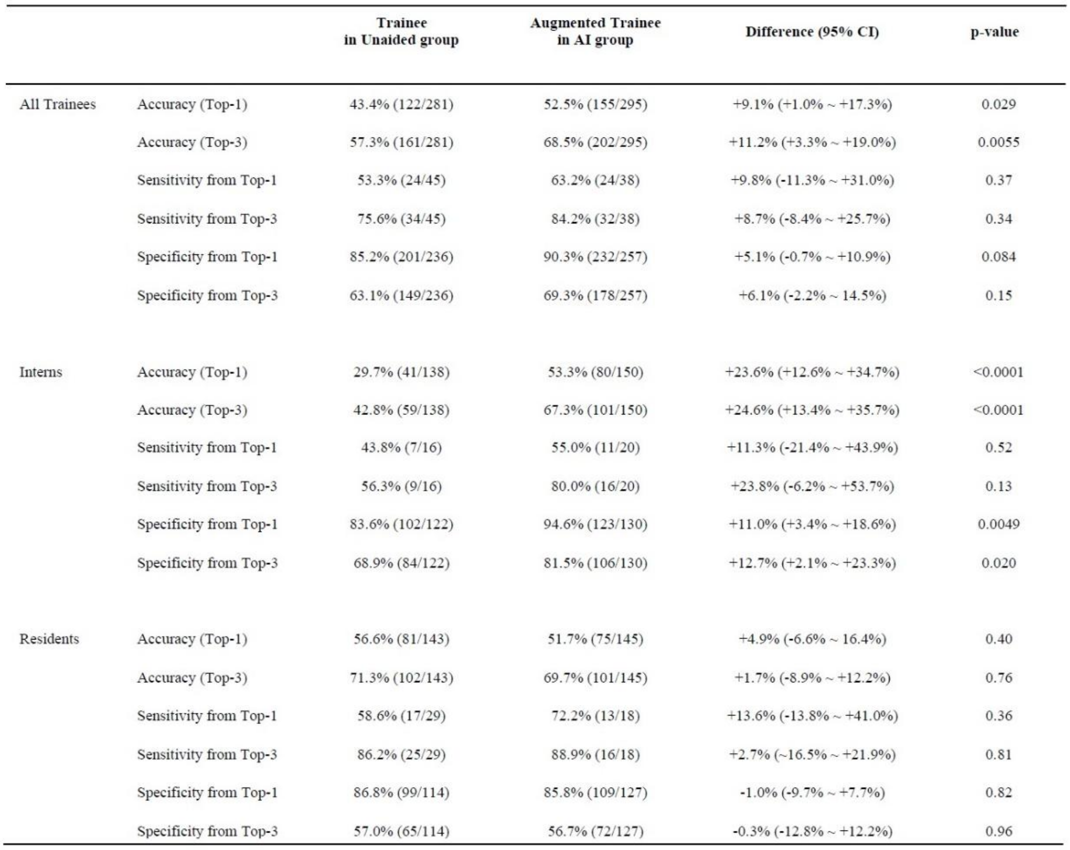
* Out-of-distribution

Supplementary Table 4. Summarized key results in the prospective before-and-after study



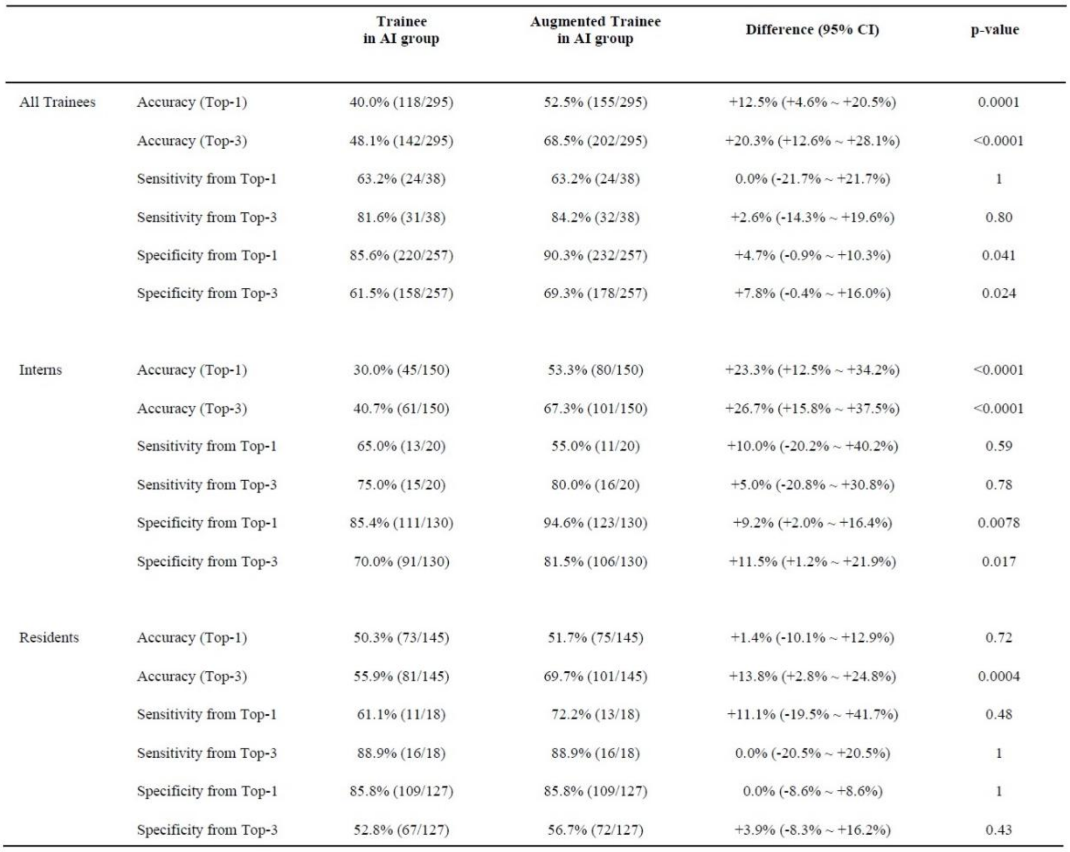
AI group	Before assistance	After assistance	*p*-value
Accuracy of the first impression			
Top-1, %	46.5	58.3	*0.008*
Top-2, %	54.2	70.1	*<0.001*
Top-3, %	54.9	71.5	*<0.001*
Determination of malignancy			
Sensitivity, %	78.3	73.9	*0.77*
Specificity, %	88.4	94.2	*0.06*
Number of differential diagnoses	1.9 ± 0.5	2.2 ± 0.6	*<0.001*
Control group			
Accuracy of the first impression			
Top-1, %	46.1	51.8	*0.19*
Top-2, %	64.5	66.7	*0.42*
Top-3, %	66.7	68.1	*0.35*
Determination of malignancy			
Sensitivity, %	65.5	65.5	*1.00*
Specificity, %	81.3	86.6	*0.09*
Number of differential diagnoses	2.0±0.4	2.1±0.5	*0.57*

Supplementary Table 5. Summary of statistics between the Unaided and AI group in the randomized controlled trial



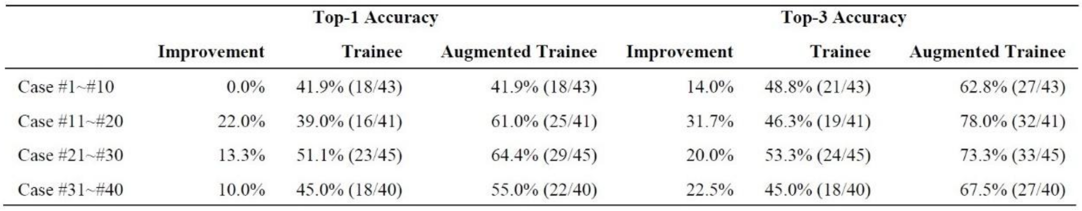
		Trainee in Unaided group	Augmented Trainee in AI group	Difference (95% CI)	p-value
All Trainees	Accuracy (Top-1)	43.4% (122/281)	52.5% (155/295)	+9.1% (+1.0% ~ +17.3%)	0.029
	Accuracy (Top-3)	57.3% (161/281)	68.5% (202/295)	+11.2% (+3.3% ~ +19.0%)	0.0055
	Sensitivity from Top-1	53.3% (24/45)	63.2% (24/38)	+9.8% (-11.3% ~ +31.0%)	0.37
	Sensitivity from Top-3	75.6% (34/45)	84.2% (32/38)	+8.7% (-8.4% ~ +25.7%)	0.34
	Specificity from Top-1	85.2% (201/236)	90.3% (232/257)	+5.1% (-0.7% ~ +10.9%)	0.084
	Specificity from Top-3	63.1% (149/236)	69.3% (178/257)	+6.1% (-2.2% ~ 14.5%)	0.15
Interns	Accuracy (Top-1)	29.7% (41/138)	53.3% (80/150)	+23.6% (+12.6% ~ +34.7%)	<0.0001
	Accuracy (Top-3)	42.8% (59/138)	67.3% (101/150)	+24.6% (+13.4% ~ +35.7%)	<0.0001
	Sensitivity from Top-1	43.8% (7/16)	55.0% (11/20)	+11.3% (-21.4% ~ +43.9%)	0.52
	Sensitivity from Top-3	56.3% (9/16)	80.0% (16/20)	+23.8% (-6.2% ~ +53.7%)	0.13
	Specificity from Top-1	83.6% (102/122)	94.6% (123/130)	+11.0% (+3.4% ~ +18.6%)	0.0049
	Specificity from Top-3	68.9% (84/122)	81.5% (106/130)	+12.7% (+2.1% ~ +23.3%)	0.020
Residents	Accuracy (Top-1)	56.6% (81/143)	51.7% (75/145)	+4.9% (-6.6% ~ 16.4%)	0.40
	Accuracy (Top-3)	71.3% (102/143)	69.7% (101/145)	+1.7% (-8.9% ~ +12.2%)	0.76
	Sensitivity from Top-1	58.6% (17/29)	72.2% (13/18)	+13.6% (-13.8% ~ +41.0%)	0.36
	Sensitivity from Top-3	86.2% (25/29)	88.9% (16/18)	+2.7% (~16.5% ~ +21.9%)	0.81
	Specificity from Top-1	86.8% (99/114)	85.8% (109/127)	-1.0% (-9.7% ~ +7.7%)	0.82
	Specificity from Top-3	57.0% (65/114)	56.7% (72/127)	-0.3% (-12.8% ~ +12.2%)	0.96

Supplementary Table 6. Summary of statistics between the before-assistance and after-assistance in the randomized controlled trial



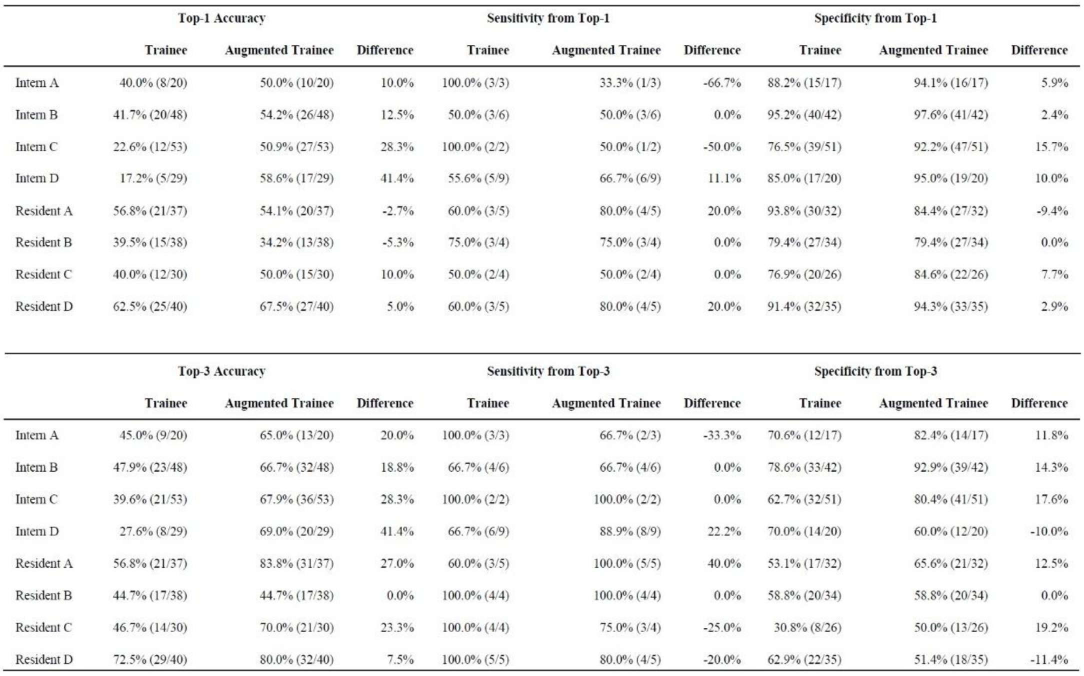
		Trainee in AI group	Augmented Trainee in AI group	Difference (95% CI)	p-value
All Trainees	Accuracy (Top-1)	40.0% (118/295)	52.5% (155/295)	+12.5% (+4.6% ~ +20.5%)	0.0001
	Accuracy (Top-3)	48.1% (142/295)	68.5% (202/295)	+20.3% (+12.6% ~ +28.1%)	<0.0001
	Sensitivity from Top-1	63.2% (24/38)	63.2% (24/38)	0.0% (-21.7% ~ +21.7%)	1
	Sensitivity from Top-3	81.6% (31/38)	84.2% (32/38)	+2.6% (-14.3% ~ +19.6%)	0.80
	Specificity from Top-1	85.6% (220/257)	90.3% (232/257)	+4.7% (-0.9% ~ +10.3%)	0.041
	Specificity from Top-3	61.5% (158/257)	69.3% (178/257)	+7.8% (-0.4% ~ +16.0%)	0.024
Interns	Accuracy (Top-1)	30.0% (45/150)	53.3% (80/150)	+23.3% (+12.5% ~ +34.2%)	<0.0001
	Accuracy (Top-3)	40.7% (61/150)	67.3% (101/150)	+26.7% (+15.8% ~ +37.5%)	<0.0001
	Sensitivity from Top-1	65.0% (13/20)	55.0% (11/20)	+10.0% (-20.2% ~ +40.2%)	0.59
	Sensitivity from Top-3	75.0% (15/20)	80.0% (16/20)	+5.0% (-20.8% ~ +30.8%)	0.78
	Specificity from Top-1	85.4% (111/130)	94.6% (123/130)	+9.2% (+2.0% ~ +16.4%)	0.0078
	Specificity from Top-3	70.0% (91/130)	81.5% (106/130)	+11.5% (+1.2% ~ +21.9%)	0.017
Residents	Accuracy (Top-1)	50.3% (73/145)	51.7% (75/145)	+1.4% (-10.1% ~ +12.9%)	0.72
	Accuracy (Top-3)	55.9% (81/145)	69.7% (101/145)	+13.8% (+2.8% ~ +24.8%)	0.0004
	Sensitivity from Top-1	61.1% (11/18)	72.2% (13/18)	+11.1% (-19.5% ~ +41.7%)	0.48
	Sensitivity from Top-3	88.9% (16/18)	88.9% (16/18)	0.0% (-20.5% ~ +20.5%)	1
	Specificity from Top-1	85.8% (109/127)	85.8% (109/127)	0.0% (-8.6% ~ +8.6%)	1
	Specificity from Top-3	52.8% (67/127)	56.7% (72/127)	+3.9% (-8.3% ~ +16.2%)	0.43

Supplementary Table 7. Accuracy changes according to the cumulative usages in the randomized controlled trial



	Top-1 Accuracy			Top-3 Accuracy		
	Improvement	Trainee	Augmented Trainee	Improvement	Trainee	Augmented Trainee
Case #1~#10	0.0%	41.9% (18/43)	41.9% (18/43)	14.0%	48.8% (21/43)	62.8% (27/43)
Case #11~#20	22.0%	39.0% (16/41)	61.0% (25/41)	31.7%	46.3% (19/41)	78.0% (32/41)
Case #21~#30	13.3%	51.1% (23/45)	64.4% (29/45)	20.0%	53.3% (24/45)	73.3% (33/45)
Case #31~#40	10.0%	45.0% (18/40)	55.0% (22/40)	22.5%	45.0% (18/40)	67.5% (27/40)

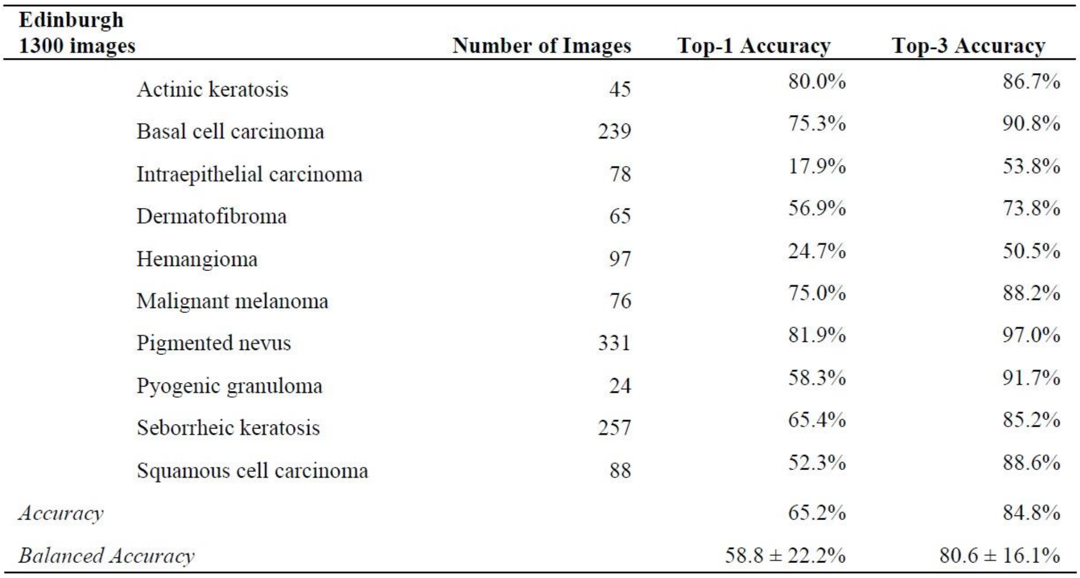
Supplementary Table 8. Accuracy, sensitivity, and specificity of each participant in the randomized controlled trial



	Top-1 Accuracy			Sensitivity from Top-1			Specificity from Top-1		
	Trainee	Augmented Trainee	Difference	Trainee	Augmented Trainee	Difference	Trainee	Augmented Trainee	Difference
Intern A	40.0% (8/20)	50.0% (10/20)	10.0%	100.0% (3/3)	33.3% (1/3)	-66.7%	88.2% (15/17)	94.1% (16/17)	5.9%
Intern B	41.7% (20/48)	54.2% (26/48)	12.5%	50.0% (3/6)	50.0% (3/6)	0.0%	95.2% (40/42)	97.6% (41/42)	2.4%
Intern C	22.6% (12/53)	50.9% (27/53)	28.3%	100.0% (2/2)	50.0% (1/2)	-50.0%	76.5% (39/51)	92.2% (47/51)	15.7%
Intern D	17.2% (5/29)	58.6% (17/29)	41.4%	55.6% (5/9)	66.7% (6/9)	11.1%	85.0% (17/20)	95.0% (19/20)	10.0%
Resident A	56.8% (21/37)	54.1% (20/37)	-2.7%	60.0% (3/5)	80.0% (4/5)	20.0%	93.8% (30/32)	84.4% (27/32)	-9.4%
Resident B	39.5% (15/38)	34.2% (13/38)	-5.3%	75.0% (3/4)	75.0% (3/4)	0.0%	79.4% (27/34)	79.4% (27/34)	0.0%
Resident C	40.0% (12/30)	50.0% (15/30)	10.0%	50.0% (2/4)	50.0% (2/4)	0.0%	76.9% (20/26)	84.6% (22/26)	7.7%
Resident D	62.5% (25/40)	67.5% (27/40)	5.0%	60.0% (3/5)	80.0% (4/5)	20.0%	91.4% (32/35)	94.3% (33/35)	2.9%

	Top-3 Accuracy			Sensitivity from Top-3			Specificity from Top-3		
	Trainee	Augmented Trainee	Difference	Trainee	Augmented Trainee	Difference	Trainee	Augmented Trainee	Difference
Intern A	45.0% (9/20)	65.0% (13/20)	20.0%	100.0% (3/3)	66.7% (2/3)	-33.3%	70.6% (12/17)	82.4% (14/17)	11.8%
Intern B	47.9% (23/48)	66.7% (32/48)	18.8%	66.7% (4/6)	66.7% (4/6)	0.0%	78.6% (33/42)	92.9% (39/42)	14.3%
Intern C	39.6% (21/53)	67.9% (36/53)	28.3%	100.0% (2/2)	100.0% (2/2)	0.0%	62.7% (32/51)	80.4% (41/51)	17.6%
Intern D	27.6% (8/29)	69.0% (20/29)	41.4%	66.7% (6/9)	88.9% (8/9)	22.2%	70.0% (14/20)	60.0% (12/20)	-10.0%
Resident A	56.8% (21/37)	83.8% (31/37)	27.0%	60.0% (3/5)	100.0% (5/5)	40.0%	53.1% (17/32)	65.6% (21/32)	12.5%
Resident B	44.7% (17/38)	44.7% (17/38)	0.0%	100.0% (4/4)	100.0% (4/4)	0.0%	58.8% (20/34)	58.8% (20/34)	0.0%
Resident C	46.7% (14/30)	70.0% (21/30)	23.3%	100.0% (4/4)	75.0% (3/4)	-25.0%	30.8% (8/26)	50.0% (13/26)	19.2%
Resident D	72.5% (29/40)	80.0% (32/40)	7.5%	100.0% (5/5)	80.0% (4/5)	-20.0%	62.9% (22/35)	51.4% (18/35)	-11.4%

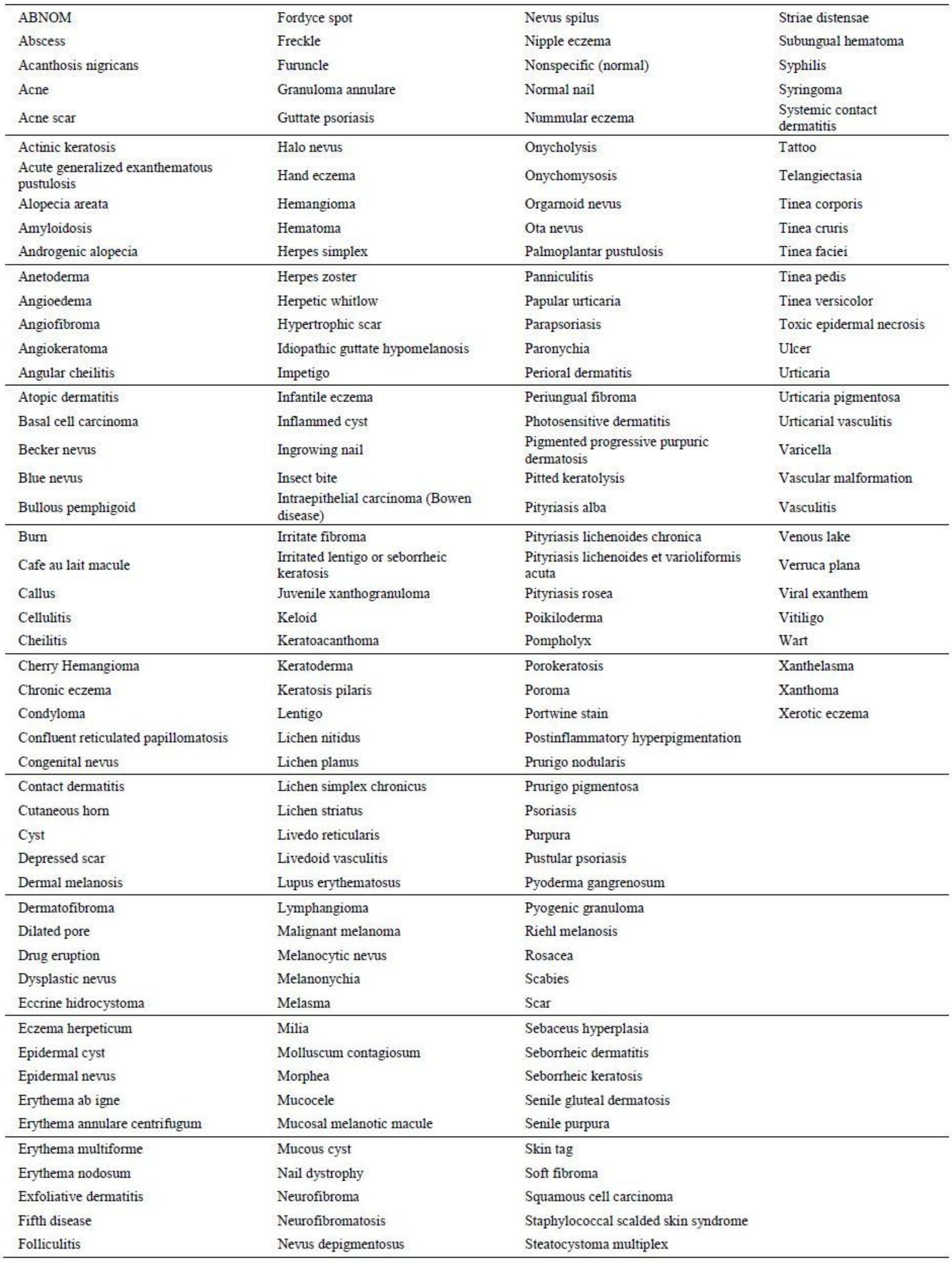
Supplementary Table 9. Top accuracies of the algorithms for 10 skin tumors in the Edinburgh Dataset



Edinburgh **1300 images**	**Number of Images**	**Top-1 Accuracy**	**Top-3 Accuracy**
Actinic keratosis	45	80.0%	86.7%
Basal cell carcinoma	239	75.3%	90.8%
Intraepithelial carcinoma	78	17.9%	53.8%
Dermatofibroma	65	56.9%	73.8%
Hemangioma	97	24.7%	50.5%
Malignant melanoma	76	75.0%	88.2%
Pigmented nevus	331	81.9%	97.0%
Pyogenic granuloma	24	58.3%	91.7%
Seborrheic keratosis	257	65.4%	85.2%
Squamous cell carcinoma	88	52.3%	88.6%
Accuracy		65.2%	84.8%
Balanced Accuracy		58.8 ± 22.2%	80.6 ± 16.1%

Using 1,300 images of the Edinburgh dataset which is commercially available for the external test (https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html), the Top-1/3 accuracies and the balanced Top-1/3 accuracies were 65.2%/84.8% and 58.8±22.2%/80.6±16.1%, respectively. The AUC for determining malignancy was 0.937 (95% CI 0.924–0.950; Edinburgh 1,300 images; Supplementary Figure 3).

Supplementary Table 10. List of 178 disorders that were trained (in-distribution disorders)



- ABNOM
- Abscess
- Acanthosis nigricans
- Acne
- Acne scar
- Actinic keratosis
- Acute generalized exanthematous pustulosis
- Alopecia areata
- Amyloidosis
- Androgenic alopecia
- Anetoderma
- Angioedema
- Angiofibroma
- Angiokeratoma
- Angular cheilitis
- Atopic dermatitis
- Basal cell carcinoma
- Becker nevus
- Blue nevus
- Bullous pemphigoid
- Burn
- Cafe au lait macule
- Callus
- Cellulitis
- Cheilitis
- Cherry Hemangioma
- Chronic eczema
- Condyloma
- Confluent reticulated papillomatosis
- Congenital nevus
- Contact dermatitis
- Cutaneous horn
- Cyst
- Depressed scar
- Dermal melanosis
- Dermatofibroma
- Dilated pore
- Drug eruption
- Dysplastic nevus
- Eccrine hidrocystoma
- Eczema herpeticum
- Epidermal cyst
- Epidermal nevus
- Erythema ab igne
- Erythema annulare centrifugum
- Erythema multiforme
- Erythema nodosum
- Exfoliative dermatitis
- Fifth disease
- Folliculitis
- Fordyce spot
- Freckle
- Furuncle
- Granuloma annulare
- Guttate psoriasis
- Halo nevus
- Hand eczema
- Hemangioma
- Hematoma
- Herpes simplex
- Herpes zoster
- Herpetic whitlow
- Hypertrophic scar
- Idiopathic guttate hypomelanosis
- Impetigo
- Infantile eczema
- Inflammed cyst
- Ingrowing nail
- Insect bite
- Intraepithelial carcinoma (Bowen disease)
- Irritate fibroma
- Irritated lentigo or seborrheic keratosis
- Juvenile xanthogranuloma
- Keloid
- Keratoacanthoma
- Keratoderma
- Keratosis pilaris
- Lentigo
- Lichen nitidus
- Lichen planus
- Lichen simplex chronicus
- Lichen striatus
- Livedo reticularis
- Livedoid vasculitis
- Lupus erythematosus
- Lymphangioma
- Malignant melanoma
- Melanocytic nevus
- Melanonychia
- Melasma
- Milia
- Molluscum contagiosum
- Morphea
- Mucocele
- Mucosal melanotic macule
- Mucous cyst
- Nail dystrophy
- Neurofibroma
- Neurofibromatosis
- Nevus depigmentosus
- Nevus spilus
- Nipple eczema
- Nonspecific (normal)
- Normal nail
- Nummular eczema
- Onycholysis
- Onychomysosis
- Organoid nevus
- Ota nevus
- Palmoplantar pustulosis
- Panniculitis
- Papular urticaria
- Parapsoriasis
- Paronychia
- Perioral dermatitis
- Periungual fibroma
- Photosensitive dermatitis
- Pigmented progressive purpuric dermatosis
- Pitted keratolysis
- Pityriasis alba
- Pityriasis lichenoides chronica
- Pityriasis lichenoides et varioliformis acuta
- Pityriasis rosea
- Poikiloderma
- Pompholyx
- Porokeratosis
- Poroma
- Portwine stain
- Postinflammatory hyperpigmentation
- Prurigo nodularis
- Prurigo pigmentosa
- Psoriasis
- Purpura
- Pustular psoriasis
- Pyoderma gangrenosum
- Pyogenic granuloma
- Riehl melanosis
- Rosacea
- Scabies
- Scar
- Sebaceus hyperplasia
- Seborrheic dermatitis
- Seborrheic keratosis
- Senile gluteal dermatosis
- Senile purpura
- Skin tag
- Soft fibroma
- Squamous cell carcinoma
- Staphylococcal scalded skin syndrome
- Steatocystoma multiplex
- Striae distensae
- Subungual hematoma
- Syphilis
- Syringoma
- Systemic contact dermatitis
- Tattoo
- Telangiectasia
- Tinea corporis
- Tinea cruris
- Tinea faciei
- Tinea pedis
- Tinea versicolor
- Toxic epidermal necrosis
- Ulcer
- Urticaria
- Urticaria pigmentosa
- Urticarial vasculitis
- Varicella
- Vascular malformation
- Vasculitis
- Venous lake
- Verruca plana
- Viral exanthem
- Vitiligo
- Wart
- Xanthelasma
- Xanthoma
- Xerotic eczema

Supplementary Methods. Algorithm in this study (Model Dermatology; https://modelderm.com)

The training history of our algorithm (Model Dermatology; https://modelderm.com) was described previously.(1-7) First, the algorithm was trained using 12 benign and malignant nodules for classification of the most common skin neoplasms.(1) As several benign disorders can mimic skin neoplasms, the algorithm should be a unified classifier that can predict a large number of disorders.(5) The ASAN and Web datasets were mainly used for training the convolutional neural networks (CNN). The ASAN dataset was assembled with 120,780 clinical images acquired from 2003 to 2016 at the Department of Dermatology at Asan Medical Center. 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. Further, as 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.(4) 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 CNN models using a transfer learning method with ImageNet pre-trained models. Histogram normalization was performed as a preprocessing step before training the models. The output values of SENet(8), SE-ResNeXt-101, SE-ResNeXt-50, ResNeSt-101(9), and ResNeSt-50 were arithmetically averaged to obtain a final model output. The hyper-parameters 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 hyper-parameters were based on previous reports.(8,10,11)

To reflect demographic metadata (age and gender), we trained a feed forward network separately. After calculating the malignancy score using the 178 outputs, were used for the input of the feed forward network. The feed-forward network consists of three inputs (malignancy score, age, and gender) as an input, three hidden layers with 200 nodes, and the last softmax layer. The feedforward network was trained using 120k images of the ASAN dataset using the NVIDIA Caffe (https://github.com/nvidia/caffe; version 0.17.2), and the hyper-parameters 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.

References for Supplementary Methods.

1. 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. *Journal of Investigative Dermatology* 2018; **138**(7): 1529-38.

2. 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. *The Journal of investigative dermatology* 2018; **138**(10): 2275-7.

3. Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality? *The Journal of investigative dermatology* 2018; **138**(10): 2277.

4. Han SS, Moon IJ, Lim W, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. *JAMA dermatology* 2020; **156**(1): 29-37.

5. Han SS, Park I, Chang SE, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology* 2020; **140**(9): 1753-61.

6. Muñoz□López C, Ramírez□Cornejo C, Marchetti MA, et al. Performance of a deep neural network in teledermatology: a single□centre prospective diagnostic study. *Journal of the European Academy of Dermatology and Venereology* 2021; **35**(2): 546-53.

7. Navarrete-Dechent C, Liopyris K, Marchetti MA. Multiclass Artificial Intelligence in Dermatology: Progress but Still Room for Improvement. *The Journal of investigative dermatology* 2021; **141**(5): 1325-8.

8. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2018; 2018. p. 7132-41.

9. Zhang H, Wu C, Zhang Z, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:200408955* 2020.

10. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; 2016. p. 770-8.

11. Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:160904836* 2016.