DR SOO ICK CHO (Orcid ID : 0000-0003-3414-9869)

DR JE-HO MUN (Orcid ID : 0000-0002-0734-2850)

PROFESSOR JIN HO CHUNG (Orcid ID : 0000-0002-0582-6392)

Article type : Original Article

Title: Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network

S. I. Cho[1*], S. Sun[2*], J. Mun[1], C. Kim[3], S. Y. Kim[3], S. Cho[4], S. W. Youn[5], H. C. Kim[6**], J. H. Chung[1**]

[1]Department of Dermatology, Seoul National University College of Medicine, Seoul, Korea

[2]Interdisciplinary Program, Bioengineering Major, Graduate School, Seoul National University, Seoul, Korea

[3]Seoul National University College of Medicine, Seoul, Korea

[4]Department of Dermatology, SMG-SNU Boramae Medical Center, Seoul, Korea

[5]Department of Dermatology, Seoul National University Bundang Hospital, Seongnam, Korea

[6]Department of Biomedical Engineering and Medical Research Center, Seoul National University College of Medicine, Seoul, Korea

*There authors contributed equally: Soo Ick Cho and Sukkyu Sun

****Co-corresponding authors:** Jin Ho Chung, MD, PhD, Department of Dermatology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul, Korea

Tel: +82-2-2072-2414

Fax: +82-2-742-7344

E-mail: jhchung@snu.ac.kr

Hee Chan Kim, PhD, Department of Biomedical Engineering and Medical Research Center, Seoul National University College of Medicine, Seoul, Korea

Tel: +82-2-741-8596

Fax: +82-2-741-8596

E-mail: hckim@snu.ac.kr

Running head: A deep convolutional neural network for malignant lip disease classification

Funding sources: None

Conflicts of Interest: None declared.

IRB approval status: Institutional Review Board of Seoul National University Hospital (IRB No. 1811-111-987)

What's already known about this topic?

- Deep convolutional neural networks (DCNNs) can classify malignant and benign skin diseases at a level equivalent to dermatologists.
- The lips are a unique feature in terms of histology, in addition to their morphology.
- Previous studies of DCNNs did not investigate tumours on specific locations.

What does this study add?

- This study shows that DCNNs can distinguish rare malignant and benign lip disorders at the same rate as dermatologists.
- DCNNs can help non-dermatologists to distinguish malignant lip diseases that are difficult to experience.

ABSTRACT

Background: Deep convolutional neural networks (DCNNs) can classify skin diseases at a level equivalent to a dermatologist, but their performance on specific areas requires further research.

Objective: We aimed to evaluate the performance of a trained DCNN-based algorithm in classifying benign and malignant lip diseases.

Methods: A training set of 1629 images (743 malignant, 886 benign) was used with Inception-Resnet-V2. Performance was evaluated using another set of 344 images and 281 images from other hospitals. Classifications by 44 participants (6 board-certified dermatologists, 12 dermatology residents, 9 medical doctors not specialized in dermatology, and 17 medical students) were used for comparison.

Results: The outcomes based on the area under curve, sensitivity, and specificity were 0.827 (95% confidence interval (CI): 0.782–0.873), 0.755 (95% CI: 0.673–0.827), and 0.803 (95% CI: 0.752–0.855), respectively for the 344 images, and 0.774 (95% CI: 0.699–0.849), 0.702 (95% CI: 0.579–0.808), and 0.759 (95% CI: 0.701–0.813), respectively for the 281 images. The DCNN was equivalent to the dermatologists and superior to the non-dermatologists when classifying malignancy. After referencing the DCNN result, the Youden's index increased significantly for the non-dermatologists, from 0.201 ± 0.156 to 0.322 ± 0.141 ($p <$ 0.001).

Conclusion: DCNNs can classify lip diseases at a level similar to dermatologists. This will help unskilled physicians discriminate between benign and malignant lip diseases.

INTRODUCTION

The lips are unique and principal feature on the face as a junctional area between keratinizing skin and oral mucosa.[1] The lips are a unique feature in terms of histology; they have a thinner cellular layer and fewer melanocytes than typical facial skin as well as no hair structures nor sweat glands. Skin lesions on the lips could be affected by biting or exposure to makeup. Thus, diseases on the lips often appear different to those on typical facial skin, which can lead to misdiagnosis.[2]

Lip cancers are relatively rare, affecting approximately 2.5 per 100,000 person-years in the United States.[3] The most common subtype of lip cancer is squamous cell carcinoma (SCC), followed by basal cell carcinoma (BCC) and malignant melanoma (MM).[4, 5] The lips are considered to be a high-risk anatomical site for skin cancer.[6, 7] Furthermore, actinic cheilitis, a precancerous lip disorder, has a risk of progressing to SCC.[8] Thus, the early diagnosis of malignant lip disorders is important for successful treatment. Unfortunately, many patients with malignant lip disorders experienced a delay to diagnosis.[9]

Deep convolutional neural networks (DCNNs) are representative deep learning models that perform better than humans in image recognition and classification.[10, 11] DCNNs have been reported that can classify dermatologic diseases at a standard equivalent to a board-certified dermatologist.[12-17] Most of these studies reported the diagnostic performance of DCNNs for skin cancer classification. However, they did not investigate tumours on specific locations such as the lips, scalp, or genital area.[18] Certain skin disorders occur exclusively or primarily on certain areas of the skin. For example, cheilitis occurs on the lips, palmoplantar pustulosis occurs on the acral area, and alopecia areata occurs on the scalp. As the number of classes increases, the accuracy of DCNNs tends to decline.[12] Therefore, the application of a DCNN algorithm specific to the area of skin can help to increase the applicability of DCNNs in dermatologic clinics.

In this study, we evaluated the performance of a DCNN in classifying malignant and benign lip disorders. In addition, we measured the changes in human decision-making before and after referencing the DCNN results.

METHODS

Datasets

The dataset in this study was taken from the photo database of Seoul National University Hospital (SNUH). The dataset contains clinical photos of dermatologic patients that were taken since October 2004. Photos related to lip diseases were selected in order to build lip datasets. The diagnoses were annotated based on the clinical and histologic findings. In addition to the SNUH photo database, clinical images of lip diseases were collected from two other affiliated hospitals, Seoul National University Bundang Hospital and SMG-SNU Boramae Medical Center. The study was approved by the Institutional Review Board of SNUH (IRB No. 1811-111-987).

Classification of lip diseases

The lip disease diagnoses were classified either malignant or benign. Malignant lip diseases included skin cancers (SCC, BCC, MM, or adenocarcinoma) and precancerous lesions (actinic cheilitis, Bowen's disease, and melanoma in situ). Benign lip diseases included lichen planus, melanotic macule, pyogenic granuloma, plasma cell cheilitis, and venous lake.

Data preparation

Blurry or images that were inadequate for clinical diagnosis were removed from the datasets. In cases where the image contained the whole face, it was cropped to show only the lip or skin lesions. In some cases, several photos were taken of same patient from multiple viewpoints. Thus, the photos were numbered based on the patient to prevent photos of the same patient from being split into both the training and testing sets.

A total of 2254 images were used for this study (Table 1). This included 1973 images of 404 patients who were diagnosed with lip disorders at SNUH between October 2004 and October 2018. In total, 34 classes of lip disorders including 4 lip cancers, 3 precancerous lip disorders, and 27 benign lip diseases were included for evaluation. The details and number of photos for each diagnosis are listed in Table 1. The images of lip diseases were classified as malignant (853 images) or benign (1120 images).

In total, 1629 of the SNUH images (743 malignant and 886 benign) were selected for the training set. The remaining 344 SNUH images (110 malignant and 234 benign) were used as the testing set along with 281 images (57 malignant and 224 benign) from Seoul National University Bundang Hospital (225 images) and SMG-SNU Boramae Medical Center (56 images).

DCNN algorithm

Various DCNN structures show good results for image classification.[19] The Inception-Resnet-V2 (Google Inc., Mountain View, CA) structure, which recorded a Top-5 error rate of 3.1% for the ILSVRC 2012 test set, was selected for this study.[20] The structure of Inception-Resnet-V2 uses factorization and residual connections. Factorization can reduce the computational power and dimensionality; hence, it can prevent overfitting, which is often problematic. Residual connections were proposed by Microsoft ResNet and they can prevent instability caused by the deeper network, which can suffer from vanishing gradients during training.[21] The hyper-parameters were set as follows: weight decay=0.00004, batch norm decay=0.9997, batch norm epsilon=0.001, dropout keep probability=0.8, and activation function ReLu with Adam-optimizer.[22]

Comparing the performance of the trained DCNN to human participants

We compared the performance of 44 medical doctors and medical students (including 6 board-certified dermatologists, 12 dermatology residents, 9 specialists who did not major in dermatology, and 17 medical students) to the trained DCNN. First, we randomly selected 40 images, 20 malignant and 20 benign. These images were viewed by the participants who were then asked: "Do you think that this lip lesion is malignant or benign?" Once the first test was finished, the participants received the output values from the trained DCNN. They were then asked the question again. The DCNN results indicated the degree of malignancy from 0 to 100.00%.

Statistical analysis

The performance of the trained DCNN was evaluated using the testing dataset of 344 images from SNUH and the 281 additional images. A receiver operating characteristic (ROC) curve was drawn in order to compare the performance of the DCNN and the participants. The area under the curve (AUC) was taken as a primary outcome while sensitivity and specificity were also calculated. Delong's test was used to compare the ROC curves generated from the two image sets. Youden's index (sensitivity + specificity – 1) was calculated before and after the DCNN results were reviewed and the values were analysed using the Wilcoxon signed rank test. Statistical analysis was performed using R statistical software version 3.51 (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Outcomes of trained DCNN on testing set

The performance outcomes of the trained DCNN based on the AUC, sensitivity, and specificity for the 344 image testing set were 0.827 (95% confidence interval (CI): 0.782–0.873), 0.755 (95% CI: 0.673–0.827), and 0.803 (95% CI: 0.752–0.855), respectively. For the additional 281 images these values were 0.774 (95% CI: 0.699–0.849), 0.702 (95% CI: 0.579–0.808), and 0.759 (95% CI: 0.701–0.813), respectively (Figure 1). There was no significant difference between the two ROC curves (p=0.229). The combined performance outcome for the 625 images in the two test tests, the AUC, sensitivity, and specificity were 0.811 (95% CI: 0.773–0.850), 0.737 (95% CI: 0.671–0.802), and 0.779 (95% CI: 0.740–0.817), respectively.

Outcomes for the trained DCNN compared to human participants

The trained DCNN outperformed most of the participants, including some board-certified dermatologists, in the ROC curve (Figure 2A).

The performance outcomes of the trained DCNN based on the AUC, sensitivity, and specificity for the 40 images chosen from the testing set were 0.792 (95% CI: 0.651–0.934), 0.850 (95% CI: 0.700–1.000), and 0.650 (95% CI: 0.450–0.850), respectively.

The sensitivity, specificity, and Youden's index for the different groups of human participants are shown in Table 2.

Outcomes of human participants changed after referencing the output value of DCNN

The sensitivity, specificity, and Youden's index for the different groups of human participants after they referenced the DCNN output are shown in Table 2. The Youden's index of the less skilled (non-dermatologist) group~~)~~ increased significantly once they referenced the DCNN output (from 0.201 ± 0.156 to 0.322 ± 0.141, $p < 0.001$), as shown in Figure 2B. However, there was no significant difference in the dermatologist group (from 0.375 ± 0.167 to 0.444 ± 0.108, $p = 0.098$).

DISCUSSION

In 2017, Esteva et al.[12] reported that DCNNs could classify skin cancer at a level equivalent to board-certified dermatologists, creating new possibilities for dermatologists. Dermatologists consider many features of skin lesions in their decision making, so such DCNN algorithms could have a significant effect on the field of dermatology.[24]

Following the article in 2017, several reports have demonstrated the application of DCNNs to dermatology. Most of these articles focused on the classification of skin cancer as the main outcome. Han et al.[13] reported that a DCNN could classify 12 benign and malignant cutaneous tumours to a standard comparable to a dermatologist. This study also showed that a DCNN algorithm mainly trained on Asian images set worked well in other testing set consisted of Caucasian skin image. Haenssle et al.[15] compared the performance of a DCNN with that of 58 dermatologists with various levels of proficiency at dermoscopy for classifying melanoma versus melanocytic nevus. The diagnostic performance of the DCNN outperformed most of the dermatologists. Tschandl et al.[17] reported that a DCNN can outperform human experts when classifying non-pigmented skin lesions. In addition, Fujisawa et al.[14] reported that a DCNN trained with 4867 clinical images outperformed dermatologists in skin tumour diagnosis. Furthermore, several studies have also reported that DCNN algorithms are equivalent to dermatologists when the skin diseases are limited to certain areas of skin, such as acral melanoma or onychomycosis discrimination.[16, 24] However, to date, there have been no studies that separate malignant and benign skin lesion on specific areas such as the lips, scalp, or genitalia.

Malignant lip disorders are rare skin diseases. Various types of cancer can occur on the lips, of which SCC is most common.[4, 5, 25] At the advanced stage, surgery for lip cancer can cause disabilities such as ingestion disorders or it may inconvenience mouth opening, which decreases the quality of life for postoperative patients.[26] MM of the lip is a rare condition and it accounts for less than 0.3% of all melanomas and all lip cancers.[27] BCC of the lip, including both cutaneous and vermilion forms, is regarded as high risk factor for recurrence.[29] Kowalski et al.[29] reported that lip carcinoma patients who experienced a provider delay of more than one month had a higher risk of reaching an advanced stage. Thus, early detection of these malignant lip disorders could improve the prognosis and outcome for patients.

Precancerous lip disorders, such as actinic cheilitis, are at risk of progressing to malignant SCC of the lip. They usually manifest as chronic inflammatory lesions in which progress slowly at older ages, hence they tend to be neglected until they reach an advanced stage.[30] Plasma cell cheilitis is a rare and benign inflammatory lip disorder. Oral involvement of lichen planus is common, although lichen planus exclusively on the lip is rare.[2] If the inflammation of the lesion is severe, benign lip diseases may be difficult to differentiate from malignant lip disorders.[31, 32]

Currently, the diagnosis of malignant lip disorders is based on biopsy and histologic findings. Prebiopsy non-invasive diagnostic approaches have been reported, including dermosopy[33, 34] and reflectance confocal microscopy.[35, 36] However, early diagnosis using these techniques is inaccessible to the general public.

In this study, the AUC-based performance for external validation (images from the two affiliated hospitals) was similar to that of the SNUH testing set. It is possible that photographs of the face are advantageous for the acquisition of images for DCNN, as no standardized measurement equipment is required. If additional regularization is required to avoid overfitting, the ability to generalize the algorithm would be limited.[37]

This study has certain limitations. First, the DCNN algorithm was used to classify dichotomous classes of lip disorders. Similarly, we collected only binary responses, and not the likelihood rating, from the participants. Second, most of the images used in this study were of Asian people. Third, the images were analysed retrospectively. Fourth, the diversity of the diagnoses from the external data of the two affiliated hospitals was lower than that of the training set. Finally, the data set was small compared to those used in previous studies. It is difficult to obtain high-quality diagnosis-annotated lip images, but these obstacles can be overcome if more appropriate images become available in the future.

In conclusion, we reported a DCNN algorithm that was trained with lip images in order to build a tool that could classify benign and malignant lip disorders. The performance of the DCNN algorithm was equivalent to that of board-certified dermatologists and it outperformed less skilled physicians. Referencing the DCNN output helped to enhance the decision-making process of less skilled participants when classifying malignant and benign lip disorders. DCNNs can help clinicians to identify malignant lip diseases early and may improve patient outcomes in the future.

ACKNOWLEDGEMENTS

We would like to thank Sungjun Choi and Hanjae Lee for their assistance in collecting the lip images. We also appreciate all the participants for their cooperation.

REFERENCES

1. Bota JP, Lyons AB, Carroll BT. Squamous cell carcinoma of the lip-A review of squamous cell carcinogenesis of the mucosal and cutaneous junction. *Dermatol Surg*. 2017; **43**:494-506.

2. Nuzzolo P, Celentano A, Bucci P, et al. Lichen planus of the lips: an intermediate disease between the skin and mucosa? Retrospective clinical study and review of the literature. *Int J Dermatol.* 2016; **55**: E473-81.

3. Canto MT, Devesa SS. Oral cavity and pharynx cancer incidence rates in the United States, 1975-1998. *Oral Oncol*. 2002; **38**:610-7.

4. Czerninski R, Zini A, Sgan-Cohen HD. Lip cancer: incidence, trends, histology and survival: 1970-2006. *Br J Dermatol*. 2010; **162**:1103-9.

5. Singer S, Zeissig SR, Emrich K, et al. Incidence of lip malignancies in Germany-data from nine population-based cancer registries. *J Oral Pathol Med*. 2017; **46**:780-5.

6. Warner CL, Cockerell CJ. The new seventh edition American Joint Committee on Cancer staging of cutaneous non-melanoma skin cancer: a critical review. *Am J Clin Dermatol.* 2011; **12**:147-54.

7. Froix AJ, Salti GI. Primary malignant melanoma of the lip. *J Surg Oncol.* 2003; **84**:7-9.

8. Jadotte YT, Schwartz RA. Solar cheilosis: an ominous precursor: part I. Diagnostic insights. *J Am Acad Dermatol*. 2012; **66**:173-84.

9. Goy J, Hall SF, Feldman-Stewart D, et al. Diagnostic delay and disease stage in head and neck cancer: a systematic review. *Laryngoscope*. 2009; **119**:889-98.

10. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015; **521**:436-44.

11. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE international conference on computer vision.* 2015; 1026-34.

12. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017; **542**:115-8.

13. Han SS, Kim MS, Lim W, et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol.* 2018; **138**:1529-38.

14. Fujisawa Y, Otomo Y, Ogata Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br J Dermatol*. 2019; **180**:373-81.

15. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018; **29**:1836-42.

16. Han SS, Park GH, Lim W, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction

of onychomycosis datasets by region-based convolutional deep neural network. *PLoS one.* 2018; **13**:e0191493.

17. Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol*. 2019; **155**:58-65.

18. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol.* 2018; **154**:1247-8.

19. Bianco S, Cadene R, Celona L, et al. Benchmark analysis of representative deep neural network architectures. IEEE Acesss 2018; **6**:64270-7.

20. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. Available at: https://arxiv.org/abs/1602.07261. (last accessed 23 August 23 2016).

21. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016; 770-8.

22. Kingma DP, Ba J. Adam: A method for stochastic optimization. Available at: https://arxiv.org/abs/1412.6980. (last accessed 30 Janunary 2017).

23. Zakhem GA, Motosko CC, Ho RS. How Should Artificial Intelligence Screen for Skin Cancer and Deliver Diagnostic Predictions to Patients? *JAMA Dermatol*. 2018; **154**:1383-4.

24. Yu C, Yang S, Kim W, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. *PLoS one*. 2018; **13**: e0193321.

25. Casal D, Carmo L, Melancia T, et al. Lip cancer: a 5-year review in a tertiary referral centre. *J Plast Reconstr Aesthet Surg.* 2010; **63**:2040-5.

26. Schuller M, Gosau M, Muller S, et al. Long-term outcome and subjective quality of life after surgical treatment of lower lip cancer. *Clin Oral Investig.* 2015; **19**:1093-9.

27. Jing G, Wu Y, Song H, et al. Primary Malignant Melanoma of the Lip: A Report of 48 Cases. *J Oral Maxillofac Surg.* 2015; **73**:2232-40.

28. Cameron MC, Lee E, Hibler BP, et al. Basal cell carcinoma Basal cell carcinoma Contemporary approaches to diagnosis, treatment, and prevention. *J Am Acad Dermatol.* 2019; **80**:321-39.

29. Kowalski LP, Franco EL, Torloni H, et al. Lateness of diagnosis of oral and oropharyngeal carcinoma: factors related to the tumour, the patient and health professionals. *Eur J Cancer B Oral Oncol.* 1994; **30B**:167-73.

30. Cavalcante AS, Anbinder AL, Carvalho YR. Actinic cheilitis: clinical and histological features. *J Oral Maxillofac Surg.* 2008; **66**:498-503.

31. Rocha N, Mota F, Horta M, et al. Plasma cell cheilitis. *J Eur Acad Dermatol Venereol.* 2004; **18**:96-8.

32. Farrier JN, Perkins CS. Plasma cell cheilitis. *Br J Oral Maxillofac Surg.* 2008; **46**:679-80.

33. Matsushita S, Kageshita T, Ishihara T. Comparison of dermoscopic and histopathological findings in a mucous melanoma of the lip. *Br J Dermatol*. 2005; **152**:1324-6.

34. Benati E, Persechino F, Piana S, et al. Dermoscopic features of squamous cell carcinoma on the lips. *Br J Dermatol.* 2017; **177**: e41-3.

35. Lupu M, Caruntu A, Caruntu C, et al. Non-invasive imaging of actinic cheilitis and squamous cell carcinoma of the lip. *Mol Clin Oncol.* 2018; **8**:640-6.

36. Maher NG, Solinas A, Scolyer RA, et al. In vivo reflectance confocal microscopy for evaluating melanoma of the lip and its differential diagnoses. *Oral Surg Oral Med Oral Pathol Oral Radiol.* 2017; **123**:84-94.

37. Yamashita R, Nishio M, Do RKG, et al. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. 2018; **9**:611-29.

Table 1. Number of photos with each lip disease diagnosis (total of 2254 images) from Seoul National University Hospital, Seoul National University Bundang Hospital, and SMG-SNU Boramae Hospital

Classification	Diagnosis	Number of photos		
		Seoul National University Hospital (n = 1973)	Seoul National University Bundang hospital (n = 225)	SMG-SNU Boramae medical center (n = 56)
Malignant (cancer)	Squamous cell carcinoma	418	13	3
	Basal cell carcinoma	7		
	Melanoma	114	2	
	Adenocarcinoma	10		
Malignant (Precancerous)	Actinic cheilitis	293	28	11
	Bowen's disease	9		
	Melanoma in situ	2		
Benign	Melanotic macule*	287	18	
	Lip dermatitis†	233	127	29
	Plasma cell cheilitis	146	18	6
	Lichen planus	124	9	7
	Venous lake	71	4	
	Pyogenic granuloma	40		
	Hemangioma	31		
	Herpes simplex	27	2	
	Mucocele	20	1	
	Fordyce's spot	20	3	
	Wart	17		
	Lupus erythematosus	13		
	Graft-versus-host disease	12		
	Hypertrophic scar	12		
	Intradermal nevus	10		
	Lymphangioma	10		
	Others‡	47		

*Including Peutz–Jeghers syndrome and Laugier–Hunziker syndrome

†Including angular cheilitis, exfoliative cheilitis, and eczematous cheilitis

‡Including angiokeratoma, aphthous ulcer, Behcet's disease, candida infection, foreign body granuloma, lichenoid keratosis, nevus flammeus, oral friction hyperkeratosis, venous malformation, vitiligo, and xanthoma

Table 2. Performance of different human participants before and after referencing the deep convolutional neural network output

	Sensitivity (mean ± standard deviation)		**Specificity** (mean ± standard deviation)		**Youden's index** (mean ± standard deviation)	
	Before	After	Before	After	Before	After
Board-certified dermatologist	0.892 ± 0.020	0.892 ± 0.092	0.575 ± 0.212	0.608 ± 0.128	0.467 ± 0.221	0.500 ± 0.100
Dermatology resident	0.796 ± 0.118	0.846 ± 0.105	0.533 ± 0.127	0.571 ± 0.123	0.329 ± 0.118	0.417 ± 0.105
Medical doctor not specialized in dermatology	0.650 ± 0.221	0.744 ± 0.151	0.461 ± 0.220	0.489 ± 0.111	0.111 ± 0.150	0.233 ± 0.148
Medical student	0.632 ± 0.150	0.832 ± 0.103	0.527 ± 0.156	0.471 ± 0.143	0.159 ± 0.129	0.303 ± 0.127

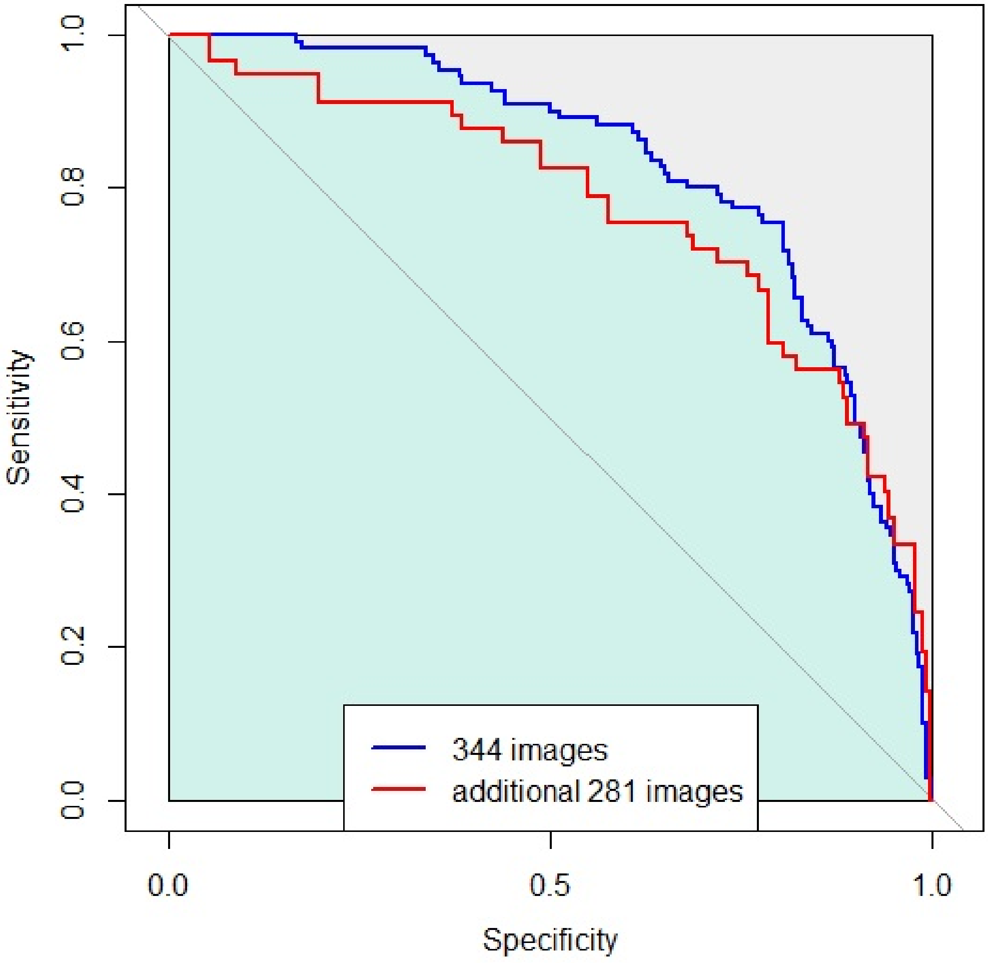
Figure legends

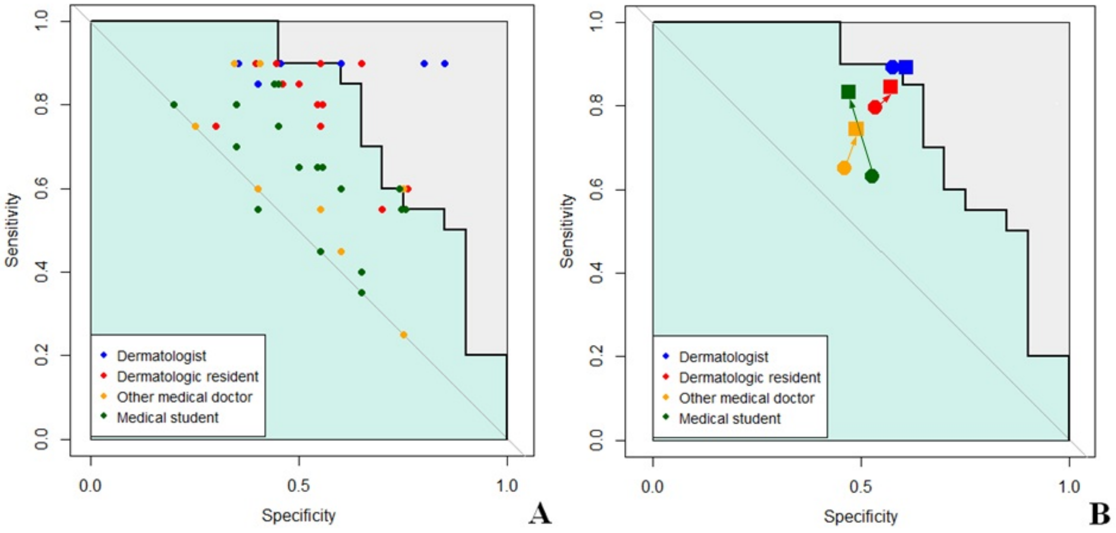
Figure 1. Receiver operating characteristic (ROC) curves for malignant lip disorders in testing set, which consisted of 344 images from Seoul National University Hospital, and 281 additional images from Seoul National University Bundang Hospital and the SMG-SNU Boramae Medical Center.

The performance outcomes of the trained DCNN based on the AUC for the 344 images in the testing set was 0.827 (95% confidence interval (CI): 0.782–0.873) and for the 281 additional 281 it was 0.774 (95% CI: 0.699–0.849). DeLong's test did not show a significant difference between the two ROC curves (*p* = 0.229).

Figure 2. Receiver operating characteristic (ROC) curves for malignant lip disorders in 40 images from the testing set.

(A) The ROC curve was drawn based on the output from the deep convolutional neural network (DCNN). The performance of the DCNN surpassed most of the participants. (B) After referencing the output of DCNN, the Youden's index of the less skilled group (participants except dermatologist) increased significantly (from 0.201 ± 0.156 to 0.322 ± 0.141, mean ± standard deviation, *p* < 0.001). The circles and squares indicate the value before and after the output was referenced, respectively.


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.5
1.0
Specificity
344 images
additional 281 images


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.5
1.0
Specificity
Dermatologist
Dermatologic resident
Other medical doctor
Medical student
A
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.5
1.0
Specificity
Dermatologist
Dermatologic resident
Other medical doctor
Medical student
B