Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm

Seung Seog Han[1,7], Myoung Shin Kim[2,7], Woohyung Lim[3], Gyeong Hun Park[4], Ilwoo Park[5] and Sung Eun Chang[6]

We tested the use of a deep learning algorithm to classify the clinical images of 12 skin diseases—basal cell carcinoma, squamous cell carcinoma, intraepithelial carcinoma, actinic keratosis, seborrheic keratosis, malignant melanoma, melanocytic nevus, lentigo, pyogenic granuloma, hemangioma, dermatofibroma, and wart. The convolutional neural network (Microsoft ResNet-152 model; Microsoft Research Asia, Beijing, China) was fine-tuned with images from the training portion of the Asan dataset, MED-NODE dataset, and atlas site images (19,398 images in total). The trained model was validated with the testing portion of the Asan, Hallym and Edinburgh datasets. With the Asan dataset, the area under the curve for the diagnosis of basal cell carcinoma, squamous cell carcinoma, intraepithelial carcinoma, and melanoma was 0.96 ± 0.01, 0.83 ± 0.01, 0.82 ± 0.02, and 0.96 ± 0.00, respectively. With the Edinburgh dataset, the area under the curve for the corresponding diseases was 0.90 ± 0.01, 0.91 ± 0.01, 0.83 ± 0.01, and 0.88 ± 0.01, respectively. With the Hallym dataset, the sensitivity for basal cell carcinoma diagnosis was 87.1% ± 6.0%. The tested algorithm performance with 480 Asan and Edinburgh images was comparable to that of 16 dermatologists. To improve the performance of convolutional neural network, additional images with a broader range of ages and ethnicities should be collected.

Journal of Investigative Dermatology (2018) **138,** 1529–1538; doi:10.1016/j.jid.2018.01.028

INTRODUCTION

Deep learning is a branch of machine learning architectures that attempts to model high-level abstractions in data using multiple processing layers. One of the deep learning models, the convolutional neural network (CNN) was used to recognize cursive numbers by LeCun in 1998 and has been shown to be useful in object recognition (Krizhevsky et al., 2012; LeCun et al., 1998). CNNs have emerged as a powerful classification tool and are consistently used in object classification competitions, including the ImageNet (http://www.image-net.org) challenge (Russakovsk et al., 2015). Since the AlexNet using a CNN architecture won the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, CNN models such as VGG, GoogLeNet, and ResNet have reported good performances in image recognition and classification (He et al., 2015; Krizhevsky et al., 2012; LeCun et al., 1998; Russakovsk et al., 2015; Simonyan and Zisserman, 2014; Szegedy et al., 2015). Microsoft ResNet (Microsoft Research Asia, Beijing, China) won the 2015 ILSVRC with an incredibly low error rate of 3.6%, significantly outperforming the human participant in the experiment, which showed that the performance of deep learning algorithms in universal object recognition and automatic speech recognition is at least on par with human ability (He et al., 2015).

Several factors have contributed the success of artificial intelligence (AI) research using neural networks, including (i) the acquisition of sufficiently large volumes of data required for the training of neural network models through the internet, (ii) improvements in graphic processing unit performance and the development of methods to use the graphic processing unit for computation, and (iii) the advancement of various deep learning methods such as rectified linear unit (i.e., ReLU), dropout, and batch normalization (Glorot et al., 2011; Ioffe and Szegedy, 2015; Srivastava et al., 2014). Despite these technological advances, however, the lack of a valid clinical dataset has limited the application of deep learning research in medicine.

Melanoma is a common skin cancer in Caucasians and has a high rate of mortality. In 2017, it was estimated that 9,730 deaths were attributable to melanoma (Siegel et al., 2017). On the other hand, basal cell carcinoma (BCC) is the most common skin cancer, and although not usually fatal, it places large burdens on health care services (Lomas et al., 2012). The development of an effective method that could discriminate skin cancer from noncancer and also classify skin cancer types would therefore be beneficial as an initial screening tool. In this study, we used a deep learning

[1]I Dermatology Clinic, Seoul, Korea; [2]Department of Dermatology, Sanggye Paik Hospital, Inje University College of Medicine, Seoul, Korea; [3]SK Telecom, Human Machine Interface Technology Laboratory, Seoul, Korea; [4]Department of Dermatology, Dongtan Sacred Heart Hospital, Hallym University College of Medicine, Dongtan, Korea; [5]Department of Radiology, Chonnam National University Medical School and Hospital, Gwangju, Korea; and [6]Department of Dermatology, Asan Medical Center, Ulsan University College of Medicine, Seoul, Korea

[7]These authors contributed equally to this work.

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

Abbreviations: AUC, area under the curve; BCC, basal cell carcinoma; CNN, convolutional neural network; ILSVRC, ImageNet Large Scale Visual Recognition Challenge

Received 16 October 2017; revised 21 January 2018; accepted 26 January 2018; accepted manuscript published online 8 February 2018; corrected proof published online 14 March 2018

algorithm (Microsoft ResNet-152) in an attempt to develop an automated classification system using the clinical images of 12 established skin disorders—BCC, squamous cell carcinoma, intraepithelial carcinoma, actinic keratosis, seborrheic keratosis, melanocytic nevus, lentigo, dermatofibroma, pyogenic granuloma, hemangioma, and wart.

RESULTS

Because dermatologists need to consider many possible impressions on a given skin image, our model was designed to list all possible candidates for a given image of the 12 types of skin disease we tested. Examples of the predictions of the ResNet-152 model for clinical images of benign and malignant tumors are shown in Figure 1. If any output of the 12 skin disorders exceeded the threshold, the model retrieved that disorder as a differential diagnosis (see Supplementary Materials and Methods online).

To improve the understanding of the prediction made by CNN and visualize the features selected by it, we implemented Grad-CAM for visual explanations from the deep network via gradient-based localization (Selvaraju et al., 2016). As shown in Figure 2, coarse and irregular portions of a lesion were determined by CNN to be important features of malignancy. This showed that the abnormal characteristics of a malignancy were learned by CNN and used as the basis for its classification of a skin malignancy (Figure 2).

The results for the area under the curve (AUC), sensitivity, and specificity of the individual disease diagnoses are listed in Table 1. In an experiment using the Asan test dataset, the AUC, sensitivity (%), and specificity (%) values for the diagnosis of BCC, squamous cell carcinoma, intraepithelial carcinoma, and melanoma were 0.96 ± 0.01, 88.8 ± 3.8, 91.7 ± 3.5; 0.83 ± 0.01, 82.0 ± 3.6, 74.3 ± 3.7; 0.82 ± 0.02, 77.7 ± 6.1, 74.9 ± 3.1; and 0.96 ± 0.00, 91.0 ± 4.3, 90.4 ± 4.5, respectively. Using the Edinburgh dataset, the algorithm slightly underperformed, producing the corresponding values of 0.90 ± 0.01, 80.1 ± 4.2, 83.0 ± 2.6; 0.91 ± 0.01, 90.2 ± 1.3, 80.0 ± 2.0; 0.83 ± 0.01, 87.2 ± 0.0, 70.5 ± 3.3; and 0.88 ± 0.01, 85.5 ± 2.3, 80.7 ± 1.1, respectively. In the case of the Hallym dataset, the sensitivity for BCC diagnosis was 87.1% ± 6.0%, with the optimal threshold setting obtained from the previous experiment using the Asan test dataset.

To differentiate between a misclassification of a malignant case as *benign* and *other malignancy,* specificities for malignant and benign cases were calculated, and their receiver operating characteristic curves were plotted, using the following equations (Figure 3):

$$\begin{aligned} &\textit{Specificity}\ (\textit{malignant}) \\ &\quad = (\textit{correctly rejected malignant conditions})/ \\ &\qquad (\textit{malignant conditions} + \textit{condition of interest}) \end{aligned}$$

$$\begin{aligned} &\textit{Specificity}\ (\textit{benign}) \\ &\quad = (\textit{correctly rejected benign conditions})/ \\ &\qquad (\textit{benign conditions} + \textit{condition of interest}). \end{aligned}$$

As depicted in Figure 3, for the four malignancies excluding melanoma from the Edinburgh dataset, specificity (benign) was higher than specificity (malignant), which indicated that the misclassification of malignant conditions as *other malignancies* was more frequent than as *benign conditions.* Actinic keratosis, which is a premalignant condition, was often misclassified as other malignancies (see Supplementary Figure S1 online). In benign conditions, the difference between specificity (benign) and specificity (malignant) was small (see Supplementary Figures S1 and S2 online).

Because of the different patient demographics in the three validation datasets we tested with our algorithm, the sensitivity and specificity of these datasets were analyzed over a change in threshold from 0.0000 to 1.0000 (Figure 4). The sensitivities of the Asan and Hallym test dataset over this threshold were similar. However, the specificities for BCC, squamous cell carcinoma, and melanoma between the Asan test dataset and Edinburgh dataset showed substantial differences, which may have been due to malignancy subtypes and the skin colors around the lesions. It may be necessary, therefore, to choose different thresholds or generate different models for different ethnic groups.

Top-1 accuracy, which is the rate at which a model yields a correct label with its top one prediction for a given image, was 57.3 ± 0.9% and 55.7 ± 1.5% for the Asan test dataset and Edinburgh dataset, respectively.

For practical purposes, 480 test images were chosen from a total set of 2,576 (1,276 Asan test images + 1,300 Edinburgh images) to compare the performances of the AI system and the dermatologists. Our AI system (Figure 3, gray curve, and see Supplementary Figures S1 and S2) showed the capability to classify 12 skin tumor types with a level of competence comparable to that of 16 dermatologists. Moreover, the AI system showed superior performance than the dermatologists in the diagnosis of BCC in the Asan test and Edinburgh datasets and in the diagnosis of melanocytic nevus from the Edinburgh dataset.

DISCUSSION

Considerable efforts continue to be made to develop automated image analysis systems for the precise detection of disease. In a previous study, computer-aided diagnostic systems relying on a feature extraction algorithm showed a promising diagnostic ability with certain skin cancers, including melanoma (Arevalo et al., 2015). However, AI with a human-engineered feature extraction could not make accurate diagnoses over a broader class of skin diseases. In recent years, deep CNNs have become very popular for use in feature learning and object classification. Extensive research from the ImageNet Large Scale Visual Recognition Challenge has indicated that the object classification capabilities of CNN architectures can even surpass those of humans (Russakovsk et al., 2015).

Several dermatologic studies have reported on the use of deep learning or machine learning (Binder et al., 1994; Codella et al., 2017; Esteva et al., 2017; Liao, 2016). Liao et al. trained a CNN to classify 23 top-level categories such as bullous disease, viral infections, and pigmented disorders using 23,000 images (Liao, 2015). The system in that study exhibited top-1 and top-5 (the rate at which a model outputs the correct label with its top one or five predictions for a given image) accuracies of 73.1% and 91.0%, respectively. In

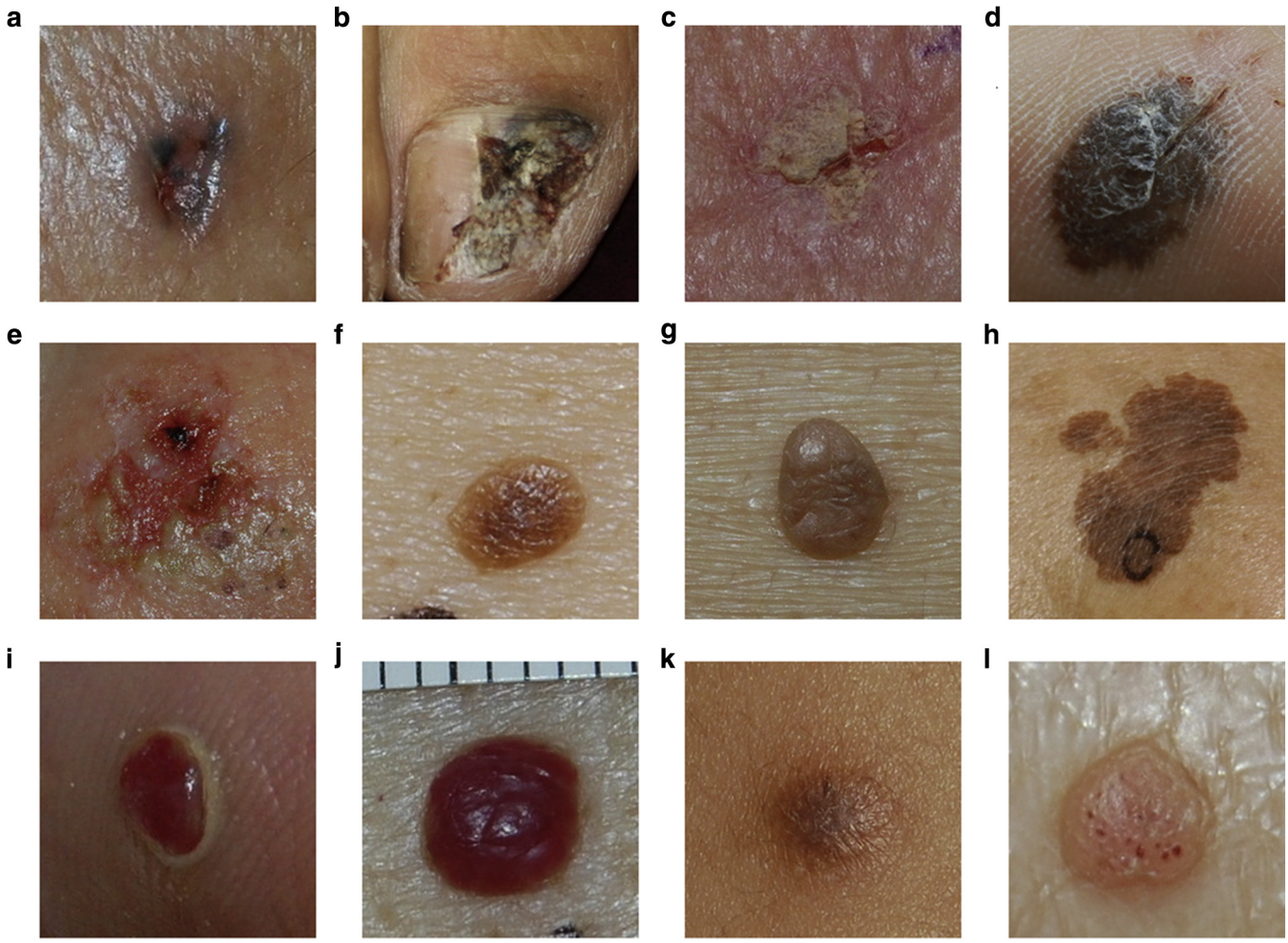

Figure 1. Representative image examples. The diagnosis for the clinical image and the output beyond the threshold were described as follows. (**a**) Basal cell carcinoma: prediction of the model = (i) basal cell carcinoma (0.9993). (**b**) Squamous cell carcinoma: prediction of the model = (i) wart (0.8893), (ii) squamous cell carcinoma (0.0473). (**c**) Intraepithelial carcinoma: prediction of the model = (i) intraepithelial carcinoma (0.9743), (ii) actinic keratosis (0.0024). (**d**) Malignant melanoma: prediction of the model = (i) malignant melanoma (0.9999). (**e**) Actinic keratosis: prediction of the model = (i) intraepithelial carcinoma (0.9348), (ii) basal cell carcinoma (0.0456), (iii) actinic keratosis (0.0127). (**f**) Seborrheic keratosis: prediction of the model = (i) melanocytic nevus (0.9209), (ii) seborrheic keratosis (0.0310), (iii) dermatofibroma (0.0173). (**g**) Melanocytic nevus: prediction of the model = (i) melanocytic nevus (0.9920), (ii) seborrheic keratosis (0.0045). (**h**) Lentigo: prediction of the model = (i) malignant melanoma (0.8799), (ii) lentigo (0.1200). (**i**) Pyogenic granuloma: prediction of the model = (i) pyogenic granuloma (0.9995). (**j**) Hemangioma: prediction of the model = (i) hemangioma (0.9927). (**k**) Dermatofibroma: prediction of the model = (i) dermatofibroma (1.0000). (**l**) Wart: prediction of the model = (i) wart (0.9918). Although the prediction by convolutional neural network for four diseases was incorrect, the differential diagnoses suggested by the algorithm are still clinically important. For example, for the image in **e**, which was diagnosed as actinic keratosis by biopsy, the observation of an erosive lesion warrants the consideration of possible malignancy and subsequent further biopsy.

a more recent study by Liao et al. (2016), the authors proposed the use of "lesion-targeted CNNs" to achieve robust skin disease diagnoses by adding lesion tags. To then exploit the correlation between skin lesions and their body site distributions, the authors built a deep multitask learning framework to jointly optimize skin lesion classification and body location classification (Liao et al., 2016, 2017). Esteva et al. (2017) have also shown dermatologist-level classification of skin cancer by deep neural networks. These researchers used 129,450 images to train and ultimately validated the system using two classes (benign/malignant). The performance of the binary (benign/malignant) classification method used by the CNN system in that report was on par with that of all of the dermatologists who participated. The authors determined an AUC of 0.96 for the diagnosis of carcinoma in 707 cases from the Edinburgh dataset and of 0.96 for the diagnosis of melanoma using 225 cases.

In this study, we aimed to classify skin lesions using the CNN in accordance with the disease category. Using the Asan test dataset, the AUC values of the ResNet-152 system for the diagnosis of BCC, squamous cell carcinoma, intraepithelial carcinoma, and melanoma were 0.96 ± 0.01, 0.83 ± 0.01, 0.82 ± 0.02, and 0.96 ± 0.00, respectively (Table 1). The corresponding AUC values using cases from the Edinburgh dataset were 0.90 ± 0.01, 0.91 ± 0.01, 0.83 ± 0.01, and 0.88 ± 0.01, respectively. With the Hallym dataset, the sensitivity for BCC diagnosis was 87.1% ± 6.0%. Because we can adjust the thresholds, the sensitivity and specificity of the CNN can be tuned to match the requirements of specific clinical settings, such as a high sensitivity for screening malignancy (Figure 4).

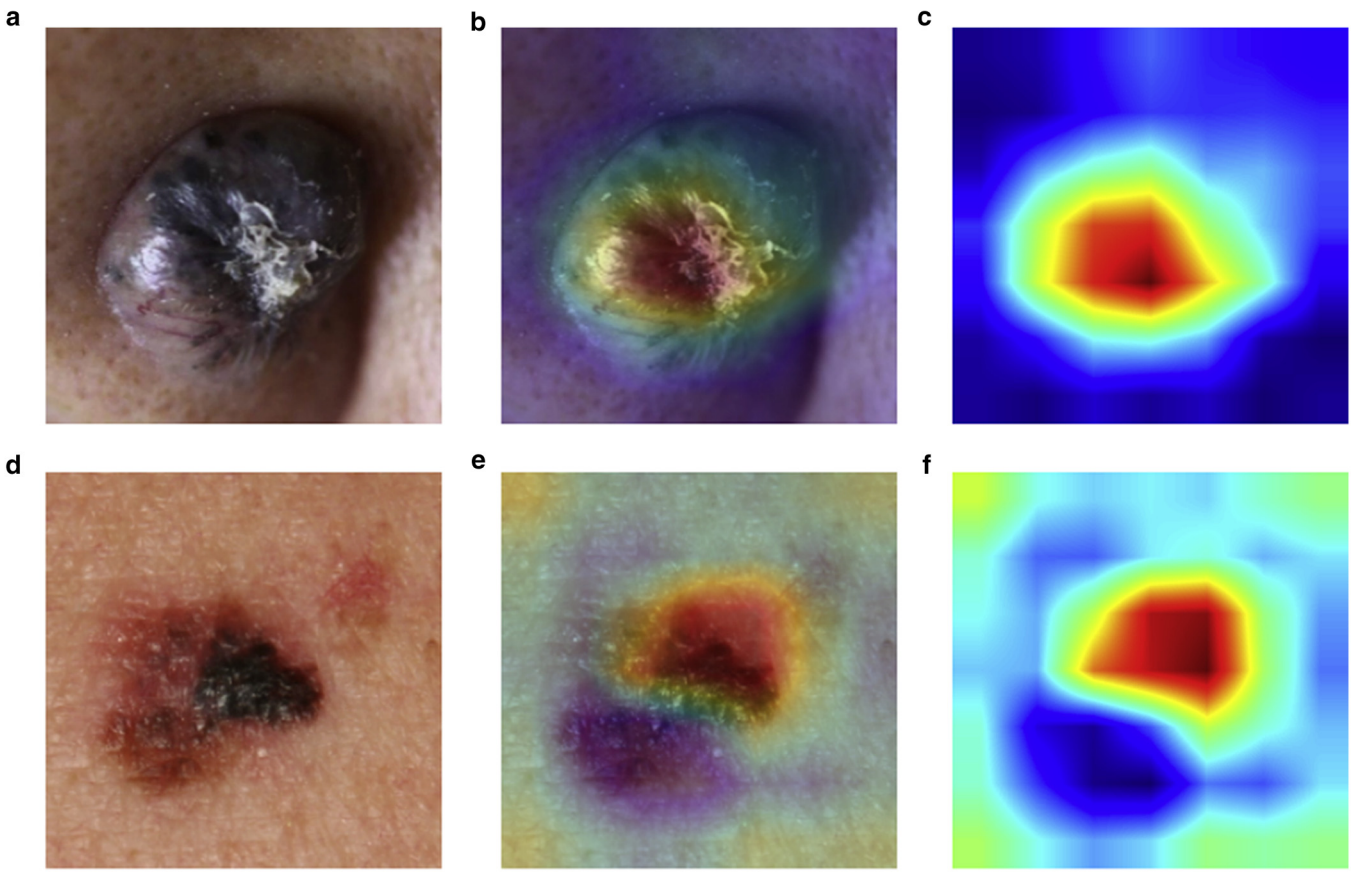

Figure 2. Visual explanations of BCC and malignant melanoma cases via gradient-based localization. (**a**) Clinical image of a BCC. (**b**) Heatmap overlaid on a clinical image of a BCC. (**c**) Heatmap for BCC. The red regions in the umbilicated central area of the lesion represent areas activated by the deep neural network. The blue background represents areas that are not activated. (**d**) Clinical image of a malignant melanoma. (**e**) Heatmap overlaid on the clinical image of a malignant melanoma. (**f**) Heatmap of malignant melanoma; the activation was focused on the dark-pigmented region with a variation in color at the upper right corner of the lesion with an irregular border. We generated a heatmap image using the python example code of Grad-CAM (https://github.com/gcucurull/CAM-Python). All of the heatmap images generated from the Asan test dataset are available at https://figshare.com/articles/Heatmap_results_of_the_Asan_test_dataset/5590900. BCC, basal cell carcinoma.

The accuracy of diagnosis with the Edinburgh dataset was slightly lower than that with the Asan dataset, possibly because of differences in patient ethnicity and the variations in overall image contrast due to unequal lighting and background. The accuracy for BCC and melanoma with the Edinburgh dataset was 0.90 ± 0.01 and 0.88 ± 0.01, respectively, compared with 0.96 ± 0.01 and 0.96 ± 0.00, respectively, with the Asan dataset (Figure 3). However, as seen in the sensitivity/threshold graph (Figure 4), the Asan and Hallym datasets, which both comprise an all-Asian patient population, generated comparable results for BCC. The clinical presentation and histological features of BCC in Asians differ from those in Caucasians. For example, approximately 75% of BCCs in Japanese patients exhibit a brown to glossy dark pigmentation, whereas only 6% of BCCs in Caucasian patients show these characteristics (Kim et al., 2009). In our pilot experiment in which we trained the model with only the Asan dataset (Asian) and validated it with Edinburgh dataset (Caucasian), the result for the BCC classification was subpar, with an AUC of 0.78 ± 0.02. To address this issue, we added the atlas dataset, consisting of 1,561 BCC images, to our training dataset and were able to obtain an improved result for BCC with an AUC of 0.90 ± 0.01. We believe that the varied subtypes in BCC among different ethnic groups were responsible for the poor results with the Edinburgh dataset in the pilot experiment and that this accuracy can be improved with the addition of datasets for various ethnic populations. In the test result of 16 dermatologists, the sensitivity for the diagnosis of BCC and melanocytic nevus cases with the Edinburgh dataset was low (< 40%), and we presume that this is because Korean dermatologists are accustomed to seeing Asian skin.

Because of its rare incidence among Asians, the number of melanomas in the Asan dataset was insufficient for our analysis, and many of the melanomas that are included were diagnosed at a late stage. In addition, the most common melanoma subtype in Asians is acral lentiginous melanoma, accounting for more than 50% of the total melanoma incidence (Kato et al., 1996; Kim et al., 2009). Consistently, the Asan dataset melanomas comprised 69.6% acral lentiginous melanomas. In our pilot study, the model trained with only the Asan dataset produced an AUC of 0.85 ± 0.01 for the classification of melanoma from the Edinburgh dataset. Because our model in the pilot study had been trained mainly with acral lentiginous melanoma, its assessment

Table 1. Summaries of the results with the Asan test dataset and Edinburgh dataset

Diagnosis	AUC	Sensitivity	Specificity	Threshold[1]
		Test = Asan Test Dataset		
Basal cell carcinoma	0.96 ± 0.01	88.8 ± 3.8	91.7 ± 3.5	0.0429 ± 0.0539
Squamous cell carcinoma	0.83 ± 0.01	82.0 ± 3.6	74.3 ± 3.7	0.0018 ± 0.0014
Intraepithelial carcinoma	0.82 ± 0.02	77.7 ± 6.1	74.9 ± 3.1	0.0030 ± 0.0024
Actinic keratosis	0.92 ± 0.01	92.5 ± 2.5	84.3 ± 2.5	0.0009 ± 0.0002
Seborrheic keratosis	0.90 ± 0.01	82.5 ± 2.1	85.6 ± 3.6	0.0172 ± 0.0140
Malignant melanoma	0.96 ± 0.00	91.0 ± 4.3	90.4 ± 4.5	0.0305 ± 0.0426
Melanocytic nevus	0.95 ± 0.01	91.5 ± 1.9	86.9 ± 1.4	0.0166 ± 0.0134
Lentigo	0.95 ± 0.01	93.9 ± 4.1	86.1 ± 2.8	0.0039 ± 0.0031
Pyogenic granuloma	0.89 ± 0.02	81.1 ± 4.7	89.6 ± 4.0	0.0014 ± 0.0019
Hemangioma	0.89 ± 0.00	81.5 ± 3.7	83.6 ± 5.2	0.1107 ± 0.0721
Dermatofibroma	0.95 ± 0.01	87.6 ± 2.8	92.6 ± 1.9	0.0227 ± 0.0191
Wart	0.94 ± 0.01	86.9 ± 2.2	86.5 ± 2.6	0.0726 ± 0.0280
Average	0.91 ± 0.01	86.4 ± 3.5	85.5 ± 3.2	0.0270 ± 0.0210
		Test = Edinburgh Dataset		
Basal cell carcinoma	0.90 ± 0.01	80.1 ± 4.2	83.0 ± 2.6	0.0996 ± 0.0771
Squamous cell carcinoma	0.91 ± 0.01	90.2 ± 1.3	80.0 ± 2.0	0.0096 ± 0.0029
Intraepithelial carcinoma	0.83 ± 0.01	87.2 ± 0.0	70.5 ± 3.3	0.0076 ± 0.0015
Actinic keratosis	0.83 ± 0.03	83.0 ± 1.3	76.5 ± 2.9	0.0003 ± 0.0002
Seborrheic keratosis	0.89 ± 0.01	79.6 ± 2.0	83.3 ± 4.3	0.0272 ± 0.0176
Malignant melanoma	0.88 ± 0.01	85.5 ± 2.3	80.7 ± 1.1	0.0032 ± 0.0010
Melanocytic nevus	0.94 ± 0.01	88.9 ± 1.1	85.4 ± 1.9	0.0332 ± 0.0144
Lentigo	—	—	—	—
Pyogenic granuloma	0.97 ± 0.01	98.6 ± 2.4	89.6 ± 1.3	0.0007 ± 0.0007
Hemangioma	0.83 ± 0.02	77.3 ± 5.2	75.7 ± 6.4	0.0238 ± 0.0194
Dermatofibroma	0.90 ± 0.00	81.0 ± 1.8	88.0 ± 3.5	0.0063 ± 0.0042
Wart	—	—	—	—
Average	0.89 ± 0.01	85.1 ± 2.2	81.3 ± 2.9	0.0212 ± 0.0139

Abbreviations: AUC, area under the curve.
[1]The optimal thresholds that maximized the sum of the sensitivity and specificity of each disorder are shown.

would most likely depend on the shape of the finger or foot to make a diagnosis. Similar to the case with BCC, the AUC for melanoma was improved to 0.88 ± 0.01 after 170 images (100 melanoma and 70 nevus) from the MED-NODE dataset and 927 images (228 melanoma, 626 nevus, and 73 lentigo) from the atlas dataset were added to the training dataset.

There have been several reports that advanced algorithms are as accurate as experienced specialists (Bejnordi et al., 2017; Esteva et al., 2017; Gulshan et al., 2016; Rajpurkar et al., 2017). The results for the Inception v3 model in diabetic retinopathy that had been trained with 128,175 retinal images were found to be similar to those for the ophthalmologists, and the diagnostic curve for the Inception v3 model in a skin malignancy study with 129,450 images was reported to be slightly better than the average for the dermatologists (Esteva et al., 2017; Gulshan et al., 2016). The ensemble model (ResNet-152 + VGG-19) achieved a superior diagnostic accuracy with onychomycosis than did 42 dermatologists by creating 49,567 standardized nail images using a region-based convolutional neural network that could detect the location of a nail in an image (Han et al., 2018). In this study, a comparable performance to our participating dermatologists was obtained by training a ResNet-152 model with over 1000 images for each of 12 skin disorders (19,398 training images in total).

Although an increased number of images for training will most likely improve the classification accuracy of a deep learning algorithm, the number of clinical images that can be collected for certain diseases may be insufficient for this purpose. The ImageNet project (http://image-net.org) is the largest visual database designed for use in visual object recognition software research. The abundant images on ImageNet have become the foundation for the future development of deep learning technology. In a similar effort, we developed a medical photographic manager, MedicalPhoto (http://medicalphoto.org), in 2007 to create a more accurate tagging method for medical images. To improve the performance of CNN in the diagnosis of skin disease, additional images with standardized diagnostics should be collected.

In summary, we investigated the possibility of establishing a classification system for 12 different cutaneous tumors using a deep learning algorithm. As technology continues to rapidly develop and as smartphones and digital cameras become increasingly accessible, the training of AI systems via deep learning technology could provide certain types of dermatological care in areas of the world where such resources are scarce. To further improve the accuracy of these systems, it will be important to increase the number of available clinical images of patients of different ages and ethnicities. This study was limited to the classification of 12

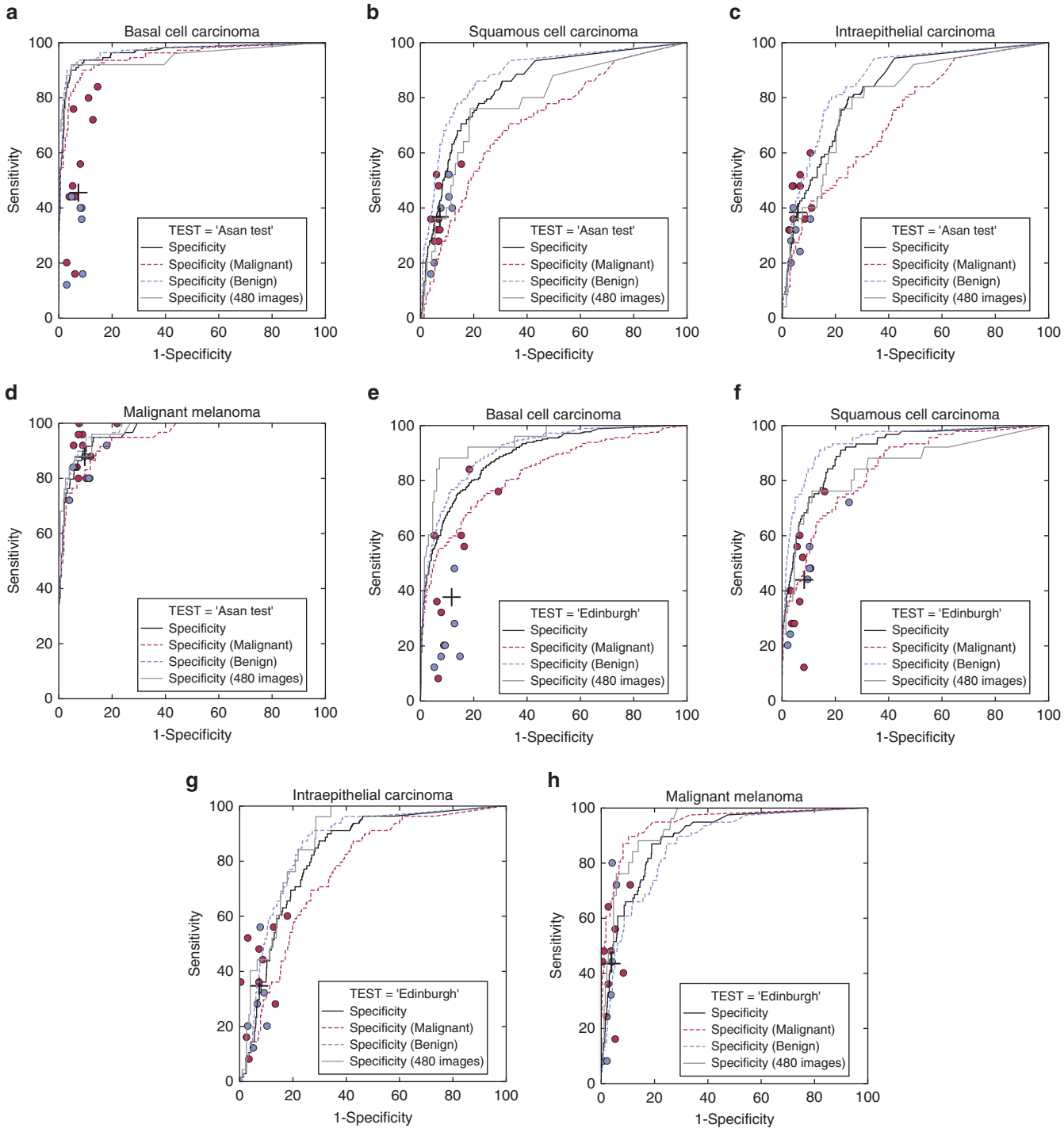

Figure 3. Receiver operating characteristic (ROC) curves for the prediction of malignancy in the Asan test dataset and Edinburgh dataset cases. ROC curves for each disorder were drawn by determining whether the output of the ResNet-152 model exceeded the threshold. ResNet-152 (gray curve) and 16 dermatologists (red dot = 10 professors; blue dot = 6 clinicians; black cross = average value of 16 dermatologists) were tested with 480 randomly chosen images from the Asan test dataset (260 images) and the Edinburgh dataset (220 images). ResNet-152 (black curve, dotted blue curve, and dotted red curve) was also tested with the whole test datasets (Asan test = 1,276 images, Edinburgh = 1,300 images). Black curve, specificity = correctly rejected conditions/all conditions. Dotted blue curve, specificity (benign) = (correctly rejected malignant conditions)/(malignant conditions + condition of interest). Dotted red curve, specificity (malignant) = (correctly rejected benign conditions)/(benign conditions + condition of interest). (**a**–**d**) Test = Asan test dataset: (**a**) BCC, (**b**) SCC, (**c**) intraepithelial carcinoma, and (**d**) Malignant melanoma. (**e**–**h**) Test = Edinburgh dataset: (**e**) BCC, (**f**) SCC, (**g**) intraepithelial carcinoma, and (**h**) malignant melanoma. BCC, basal cell carcinoma; SCC, squamous cell carcinoma.

skin disorders, and it will necessary to expand the repertoire of skin images to include other cutaneous tumors and normal skin types and thereby reduce the false-positive rate when using deep learning algorithms in real clinical practice.

MATERIALS AND METHODS

Data Collection

Data on patient demographics and clinical images were collected via a retrospective chart review, and patient consent was not

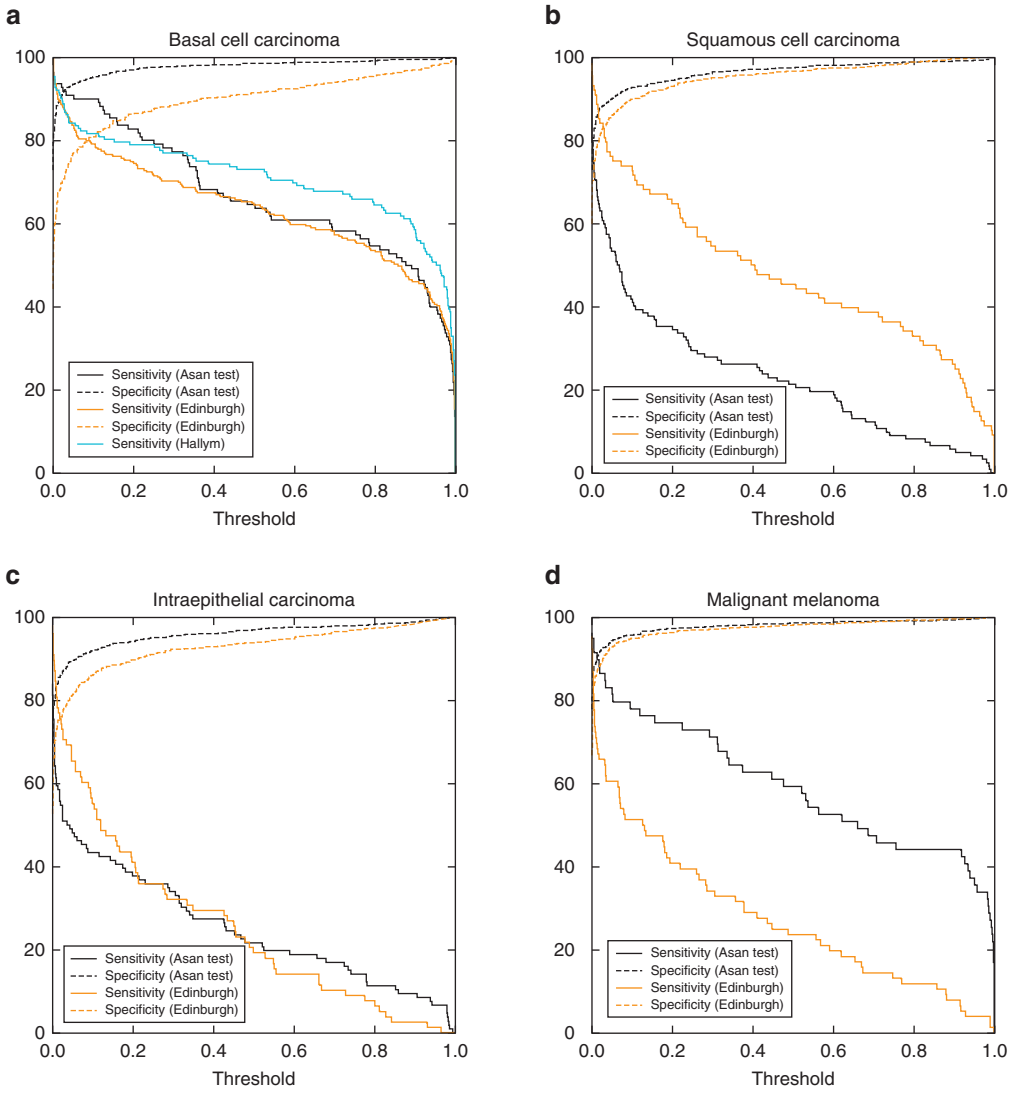

Figure 4. Threshold/sensitivity/specificity graphs for the diagnosis of four malignancies with various datasets. In general, as the threshold increases, sensitivity decreases and specificity increases. However, because of differences between the subtypes of malignancies among different races, differences in skin color around the lesion, ranges in age and sex, and different lighting conditions from various cameras, the change in the sensitivity and specificity over the threshold increase varies in each dataset. The Hallym dataset consisted of BCC cases alone and thus only the sensitivity for BCC was plotted. (**a**) BCC. (**b**) SCC. (**c**) Intraepithelial carcinoma. (**d**) Malignant melanoma. BCC, basal cell carcinoma; SCC, squamous cell carcinoma.

necessary (Asan Medical Center institutional review board approval number 2017-0087). The first dataset, known as the *Asan dataset,* was collected from the Department of Dermatology at Asan Medical Center. A total of 598,854 clinical images taken from 2000 through 2016 were considered for use in this study. For the accurate and effective diagnosis of skin disease and the efficient handling of a large dataset, we developed a noncommercial, open-source medical image management tool, MedicalPhoto (http://medicalphoto.org). The standardized ICD-10 diagnosis and clinical impression has been attached to each image since 2007 using MedicalPhoto. In addition, the pathological findings for all images from patients who underwent a biopsy were recorded.

From the Asan dataset, 12 types of skin disease were selected based on the following criteria. Because a large number of images are needed for deep learning model training, we searched for disorders with more than 1,000 images in the entire Asan dataset. After excluding postoperative and inadequate images, we finally selected 17,125 images (12,656 of which were confirmed by biopsy) for 12 disorders from 4,867 patients who had benign or malignant tumors (Table 2).

Because of large differences in the subtypes of BCC and malignant melanoma between Asians and Caucasians (Kato et al., 1996; Kim et al., 2009), the subtypes of these lesions were distinguished by examining each image of the Asan dataset. The BCC cases consisted of the following subtypes: nodular, 882 images (81.5%); superficial, 117 images (10.8%); and infiltrative, 87 images (8.0%). Among all of the BCC images, 884 (81.7%) involved pigmentation (pigmented BCC). Among the melanoma images, 417 out of 599 (69.6%) were acral lentiginous melanoma.

In addition to the 12 skin diseases used for the classification, 248 diseases from 159,477 images among 17,888 Asan Medical Center patients were prepared to train the CNN model. If the CNN model was trained with only 12 types of skin diseases, then the CNN model classify one of these disorders even when given an image that does not belong to one of 12 disorders. The model training with the additional datasets, known as the *additional Asan dataset,* improves the specificity of this disease classification.

The second dataset used, the *MED-NODE dataset* (http://www.cs.rug.nl/~imaging/databases/melanoma_naevi), contained 70 melanoma and 100 nevus images from the Department of Dermatology at the University Medical Center Groningen (Giotis et al., 2015).

Table 2. Summary of the image characteristics and demographic information in the Asan, Hallym, MED-NODE, atlas, and Edinburgh datasets

Characteristics	Asan[1]	Additional Asan[2]	Atlas[3]
Images, n	17,125 (12,656)	159,477	3,820
Patient demographics			
Unique individuals, n	4,867 (3,952)	17,888	—
Age in years, mean ± SD	47.37 ± 22.91 (50.92 ± 20.55)	40.91 ± 20.68	—
Male, %	44.7 (45.5)	43.4%	—
Race	>99% Asian	>99% Asian	Mainly Caucasian
Diagnosis, number of images			
Basal cell carcinoma	1,082 (1,082)	—	1,561
Squamous cell carcinoma	1,231 (1,231)	—	—
Intraepithelial carcinoma	918 (918)	—	—
Actinic keratosis	651 (631)	—	—
Seborrheic keratosis	1,423 (995)	—	897
Malignant melanoma	599 (599)	—	228
Melanocytic nevus	2,706 (2,366)	—	626
Lentigo	1,193 (490)	—	73
Pyogenic granuloma	375 (365)	—	—
Hemangioma[4]	2,715 (832)	—	—
Dermatofibroma	1,247 (1,153)	—	—
Wart	2,985 (1,994)	—	435
Others	—	159,477	—
	MED-NODE[3]	**Hallym**	**Edinburgh[3]**
Images, n	170	152	1300
Patient demographics			
Unique individuals, n	—	106	—
Age in years, mean ± SD	—	67.6 ± 13.3	—
Male, %	—	48.1	—
Race	Mainly Caucasian	104 Asian, 2 Caucasian	Mainly Caucasian
Diagnosis, number of images			
Basal cell carcinoma	—	152	239
Squamous cell carcinoma	—	—	88
Intraepithelial carcinoma	—	—	78
Actinic keratosis	—	—	45
Seborrheic keratosis	—	—	257
Malignant melanoma	100	—	76
Melanocytic nevus	70	—	331
Lentigo	—	—	—
Pyogenic granuloma	—	—	24
Hemangioma	—	—	97
Dermatofibroma	—	—	65
Wart	—	—	—
Others	—	—	—

Abbreviation: SD, standard deviation.

[1]The number of images confirmed via biopsy is indicated in parentheses. All of the test sets came from repositories of biopsy-proven images. The Asan dataset was partitioned into training and testing datasets as follows: Asan training dataset = 90% of the Asan dataset used for training, Asan test dataset = 10% of the Asan dataset used for testing. Thumbnails of the images from the Asan and Hallym datasets are available at https://figshare.com/articles/Asan_and_Hallym_Dataset_Thumbnails_/5406136.

[2]Additional Asan dataset consisted of 248 diseases, and unique patient codes and individual data, including age and sex, were available for 99.6% of the images.

[3]No demographic data were available from the Atlas, MED-NODE, and Edinburgh datasets.

[4]From the image findings, cherry angioma (senile angioma) represented 65% of the Edinburgh dataset, and there were no cases of infantile hemangioma. In the Asan test dataset, cherry angioma represented 30% and other vascular tumors 70% of the images, also with no cases of infantile hemangioma.

The third dataset, known as the *atlas dataset,* was obtained from several dermatologic atlas sites (http://dermquest.com, http://www.dermatlas.net, http://www.dermis.net/dermisroot/en/home/index.htm, http://www.meddean.luc.edu/lumen/MedEd/medicine/dermatology/melton/atlas.htm, http://www.dermatoweb.net, http://www.danderm-pdv.is.kkh.dk/atlas/index.html, http://www.atlasdermatologico.com.br, http://www.hellenicdermatlas.com/en). A total of 3,820 images were downloaded from these websites and used as a training dataset.

The fourth dataset, the *Edinburgh dataset,* was obtained from the Edinburgh Dermofit Image Library (https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html), and consisted of 1,300 images of 10 disorders to be used for model validation.

The final dataset, known as the *Hallym dataset,* consisted of 152 BCC images obtained from 106 patients treated between 2010 and 2016 at Dongtan Sacred Heart Hospital, Hallym University, and Sanggye Paik Hospital, Inje University, and was also used for the CNN model validation.

Images from 12 benign and malignant skin tumors from the Asan dataset were used as a training dataset for our deep learning algorithm. After the images from the Asan dataset were sorted by time, the oldest 90% (15,408 images) were used as a training dataset (*Asan training dataset*) and the remainder (1,276 images) as a test dataset (*Asan test dataset*). Along with the Asan training dataset, the additional Asan, MED-NODE, and atlas datasets were used to train the model. The Edinburgh, Hallym, and Asan test datasets were used for validation. All of these test sets came from repositories of biopsy-proven images.

Because the Asan dataset consisted of images from an Asian population, inclusion of the MED-NODE and atlas datasets helped diversify the ethnicity of the skin images used.

There are several software tools now available for deep learning, including Caffe (http://caffe.berkeleyvision.org; Berkeley Vision and Learning Center, Berkeley, CA), Torch (http://torch.ch), MXNet (https://mxnet.apache.org; Apache Software Foundation, Forest Hill, MD), Microsoft cognitive toolkit (CNTK; https://github.com/Microsoft/CNTK; Microsoft Research AI, Redmond, WA), and Tensorflow (https://www.tensorflow.org; Google Brain Team, Mountain View, CA); in this study, we used Caffe. We chose Microsoft ResNet-152 as our CNN model because it is, along with Google Inception, currently one of the cutting-edge CNN models (Canziani et al., 2016). We fine-tuned the ImageNet pretrained model of the ResNet-152. The lesion of interest was cropped and resized to a low resolution (224 × 224 pixels) for training and testing. Detailed image preprocessing and training parameters are available in the Supplementary Materials and Methods.

The best sensitivity, specificity, and the optimal threshold that maximized Youden index (sensitivity + specificity − 1) were determined. We drew a receiver operating characteristic curve by judging whether the outputs of the ResNet-152 model exceeded the threshold or not (see Supplementary Materials and Methods).

We compared the performances of the panel of dermatologists and our algorithm using biopsy-proven images from the Asan test dataset and Edinburgh dataset. We randomly selected 25 images of each malignancy and 20 images of each benign lesion (260 images from Asan test, 220 images from Edinburgh). Sixteen dermatologist board members (10 professors and 6 clinicians, each with over 10 years of experience) viewed the original-resolution photograph as a PDF file and was asked, "What is your diagnosis for the following image among the 12 diagnoses?" (multiple choice). We informed these clinicians of the fact that the test images were from a random mixture of Asian and Caucasian patients. The test PDF document from the Asan dataset and the image identifications from the Edinburgh dataset can be found at https://figshare.com/articles/AI_vs_Dermatologist_Test_PDF_Docoment/5592631.

We created a user-friendly PC- and smartphone-based automatic skin disease classification test platform using the malignancy model described herein so that our algorithm can be tested by any interested clinicians or researchers around the world (http://dx.medicalphoto.org). The Caffe model files created in this study are available at https://figshare.com/articles/Caffemodel_files_and_Python_Examples/5406223.

ORCIDs
Seung Seog Han: http://orcid.org/0000-0002-0500-3628
Sung Eun Chang: http://orcid.org/0000-0003-4225-0414

CONFLICT OF INTEREST
LW is employed by SK Telecom. However, the company did not have any role in the study design, data collection and analysis, decision to publish, or preparation of this manuscript. This study is not related to any research being conducted by SK Telecom.

ACKNOWLEDGMENTS
We are indebted to Eun Jin Yeon and Ik Jun Moon for their assistance in writing this manuscript.

SUPPLEMENTARY MATERIAL
Supplementary material is linked to the online version of the paper at www.jidonline.org, and at https://doi.org/10.1016/j.jid.2018.01.028.

REFERENCES
Arevalo J, Cruz-Roa A, Arias V, Romero E, Gonzalez FA. An unsupervised feature learning framework for basal cell carcinoma image analysis. Artif Intell Med 2015;64:131–45.

Bejnordi BE, Veta M, van Diest PJ, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199–210.

Binder M, Steiner A, Schwarz M, Knollmayer S, Wolff K, Pehamberger H. Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. Br J Dermatol 1994;13:460–5.

Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications, https://arxiv.org/abs/1605.07678; 2016 (accessed 14 April 2017).

Codella N, Nguyen Q-B, Pankanti S, Gutman D, Helba B, Halpern A, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J Res Dev 2017;61(4).

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.

Giotis I, Molders N, Land S, Biehl M, Jonkman MF, Petkov N. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst Appl 2015;42:6578–85.

Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. J Machine Learn Res 2011;15:315–23.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.

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.

He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, https://arxiv.org/abs/1502.01852; 2015 (accessed 6 February 2015).

Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift, https://arxiv.org/pdf/1502.03167v3.pdf; 2015 (accessed 6 February 2015).

Kato T, Suetake T, Sugiyama Y, Tabata N, Tagami H. Epidemiology and prognosis of subungual melanoma in 34 Japanese patients. Br J Dermatol 1996;134:383–7.

Kim GK, Del Rosso JQ, Bellew S. Skin cancer in asians: part 1: nonmelanoma skin cancer. J Clin Aesthet Dermatol 2009;2(8):39–42.

Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012;25:1097–105.

LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86:2278–324.

Liao H. A deep learning approach to universal skin disease classification. University of Rochester Department of Computer Science, CSC 400 graduate problem seminar, project report. https://www.semanticscholar.org/paper/A-Deep-Learning-Approach-to-Universal-Skin-Disease-Liao/af34fc0aebff011b56ede8f46ca0787cfb1324ac; 2015 (accessed 18 February 2018).

Liao H, Li Y, Luo J. Skin disease classification versus skin lesion characterization: achieving robust diagnosis using multi-label deep neural networks. In: Pattern Recognition (ICPR), 2016 23rd International Conference on IEEE;2016:355-360.

Liao H, Li J. A deep multitask learning approach to skin lesion classification. Presented at: AAAI 2017 Joint Workshop on Health Intelligence. 4–5 February 2017; San Francisco, CA.

Lomas A, Leonardi-Bee J, Bath-Hextall F. A systematic review of worldwide incidence of nonmelanoma skin cancer. Br J Dermatol 2012;166:1069–80.

Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning, https://arxiv.org/abs/1711.05225; 2017 (accessed 25 December 2017).

Russakovsk O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211–52.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: why did you say that? visual explanations from deep networks via gradient-based localization, https://arxiv.org/abs/1610.02391; 2016 (accessed 21 March 2017).

Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RG, Barzi A, et al. Colorectal cancer statistics, 2017. CA Cancer J Clin 2017;67:177–93.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition, https://arxiv.org/abs/1409.1556; 2014 (accessed 23 December 2014).

Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions, https://arxiv.org/abs/1409.4842; 2015 (accessed 17 September 2014).