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) ■, ■–■; 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 XXX

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