LETTER

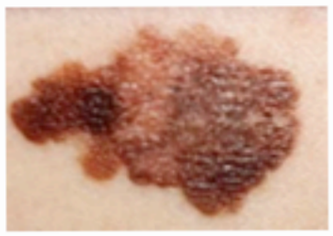
doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

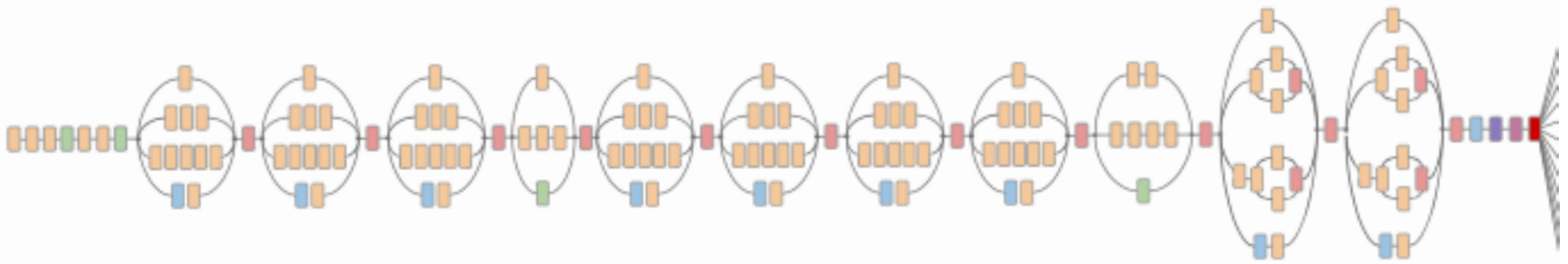
Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Skin cancer, the most common human malignancy[1–3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4,5] show potential for general and highly variable tasks across many fine-grained object categories[6–11]. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets[12]—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the identification of the most common cancers, the second represents the identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 (ref. 13) and can therefore potentially provide low-cost universal access to vital diagnostic care.

Skin lesion image



Deep convolutional neural network (Inception v3)



- Convolution
- AvgPool
- MaxPool
- Concat
- Dropout
- Fully connected
- Softmax

Training classes (757)

- Acral-lentiginous melanoma
- Amelanotic melanoma
- Lentigo melanoma
- ...
- Blue nevus
- Halo nevus
- Mongolian spot
- ...

Inference classes (varies by task)

- 92% malignant melanocytic lesion
- 8% benign melanocytic lesion

Figure 1 | Deep CNN layout. Our classification technique is a deep CNN. Data flow is from left to right: an image of a skin lesion (for example, melanoma) is sequentially warped into a probability distribution over clinical classes of skin disease using Google Inception v3 CNN architecture pretrained on the ImageNet dataset (1.28 million images over 1,000 generic object classes) and fine-tuned on our own dataset of 129,450 skin lesions comprising 2,032 different diseases. The 757 training classes are defined using a novel taxonomy of skin disease and a partitioning algorithm that maps diseases into training classes

(for example, acrolentiginous melanoma, amelanotic melanoma, lentigo melanoma). Inference classes are more general and are composed of one or more training classes (for example, malignant melanocytic lesions—the class of melanomas). The probability of an inference class is calculated by summing the probabilities of the training classes according to taxonomy structure (see Methods). Inception v3 CNN architecture reprinted from https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html

lesions. In this task, the CNN achieves $72.1 \pm 0.9\%$ (mean $\pm$ s.d.) overall accuracy (the average of individual inference class accuracies) and two dermatologists attain 65.56% and 66.0% accuracy on a subset of the validation set. Second, we validate the algorithm using a nine-class disease partition—the second-level nodes—so that the diseases of each class have similar medical treatment plans. The CNN achieves $55.4 \pm 1.7\%$ overall accuracy whereas the same two dermatologists attain 53.3% and 55.0% accuracy. A CNN trained on a finer disease partition performs better than one trained directly on three or nine classes (see Extended Data Table 2), demonstrating the effectiveness of our partitioning algorithm. Because images of the validation set are labelled by dermatologists, but not necessarily confirmed by biopsy, this metric is inconclusive, and instead shows that the CNN is learning relevant information.

Data availability statement. The medical test sets that support the findings of this study are available from the ISIC Archive (https://isic-archive.com/) and the Edinburgh Dermofit Library (https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html). Restrictions apply to the availability of the medical training/ validation data, which were used with permission for the current study, and so are not publicly available. Some data may be available from the authors upon reasonable request and with permission of the Stanford Hospital.

Data availability statement. The medical test sets that support the findings of this study are available from the ISIC Archive (https://isic-archive.com/) and the Edinburgh Dermofit Library (https://licensing.eri.ed.ac.uk/i/software/dermofit- image-library.html). Restrictions apply to the availability of the medical training/ validation data, which were used with permission for the current study, and so are not publicly available. Some data may be available from the authors upon reasonable request and with permission of the Stanford Hospital.

Figure 2 | A schematic illustration of the taxonomy and example test set images. a, A subset of the top of the tree-structured taxonomy of skin disease. The full taxonomy contains 2,032 diseases and is organized based on visual and clinical similarity of diseases. Red indicates malignant, green indicates benign, and orange indicates conditions that can be either. Black indicates melanoma. The first two levels of the taxonomy are used in validation. Testing is restricted to the tasks of **b**. **b**, Malignant and benign

DERMOFIT IMAGE LIBRARY

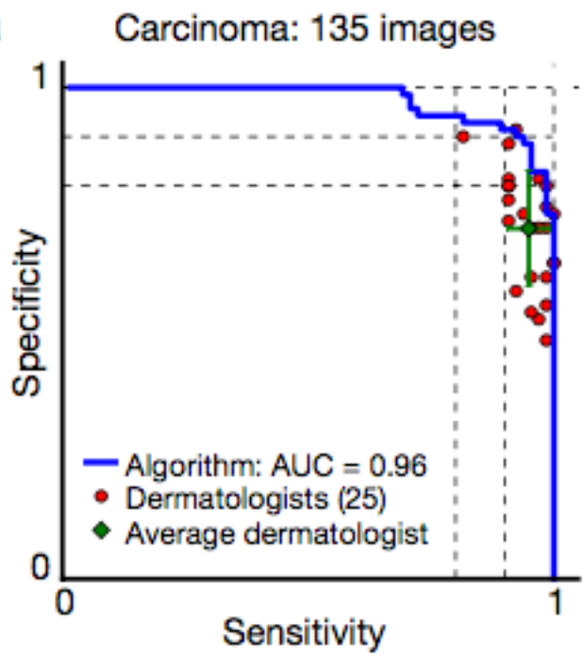
High quality skin lesion images for use as a research tool in computer science and medical imaging

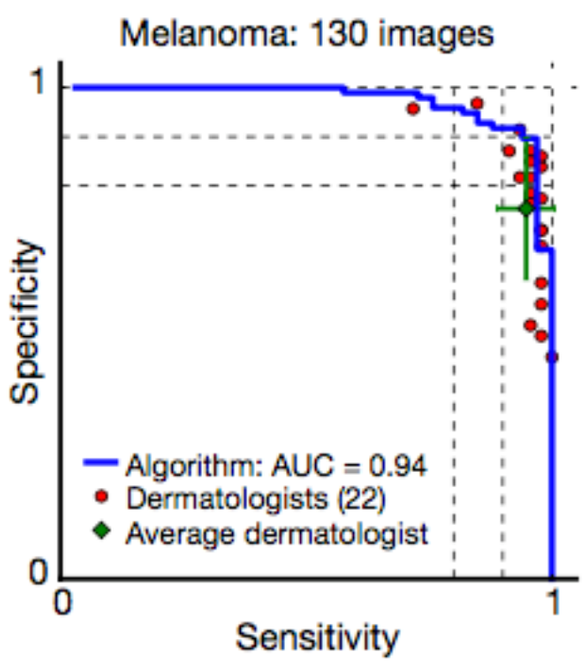
OPTIONS

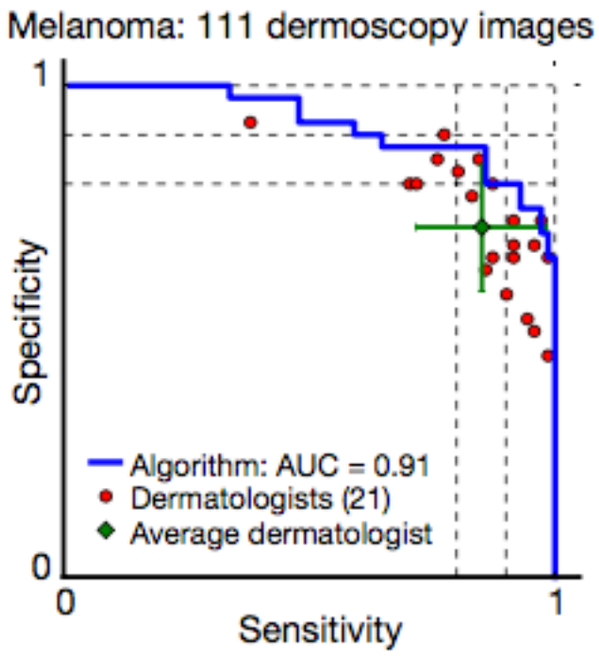
Please check carefully that the terms you select correspond to your intended use of the product.

ACADEMIC LICENCE AGREEMENT FOR DERMOFIT IMAGE LIBRARY 4.00

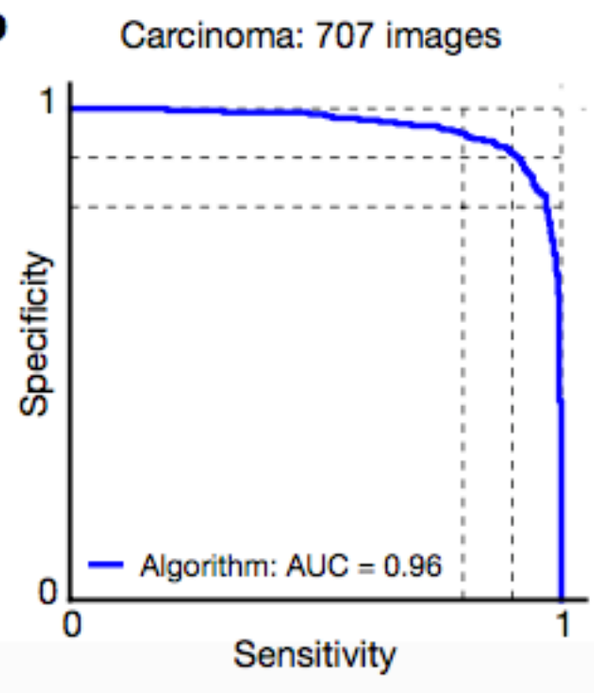
a

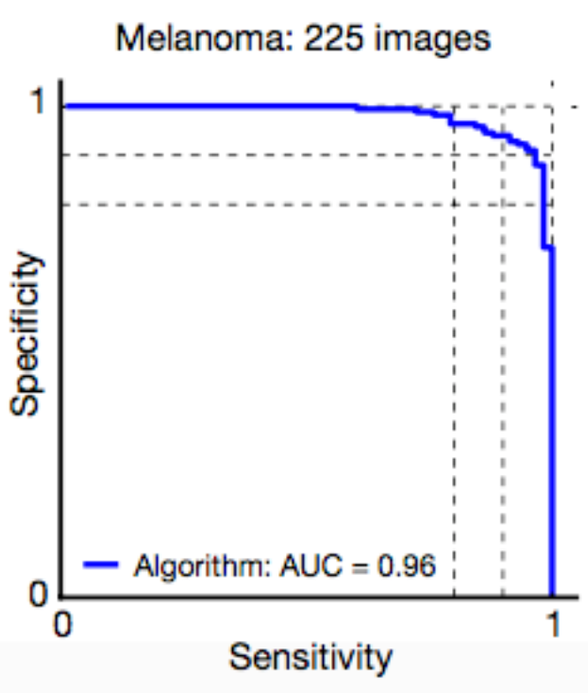

Carcinoma: 135 images
Specificity
Sensitivity
1
0
0
1
Algorithm: AUC = 0.96
Dermatologists (25)
Average dermatologist

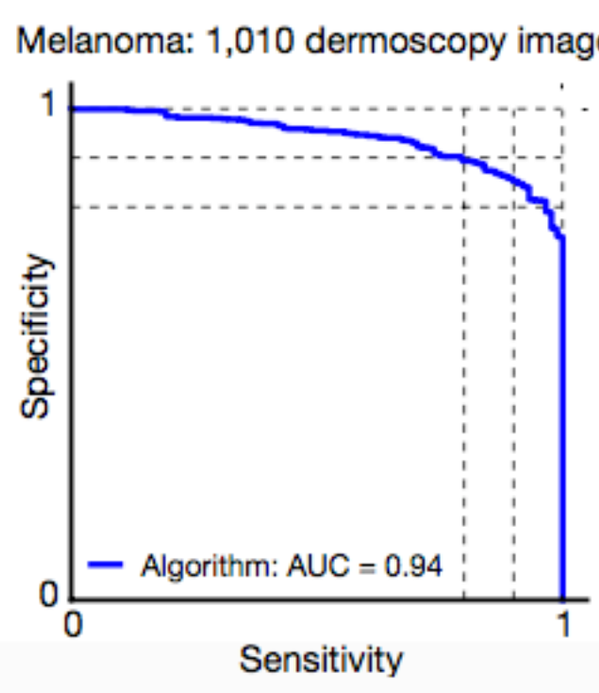

Melanoma: 130 images
Specificity
Sensitivity
1
0
0
1
Algorithm: AUC = 0.94
Dermatologists (22)
Average dermatologist


Melanoma: 111 dermoscopy images
Specificity
Sensitivity
1
0
0
1
Algorithm: AUC = 0.91
Dermatologists (21)
Average dermatologist

b


Carcinoma: 707 images
Specificity
Sensitivity
1
0
0
1
Algorithm: AUC = 0.96


Melanoma: 225 images
Specificity
Sensitivity
1
0
0
1
Algorithm: AUC = 0.96


Melanoma: 1,010 dermoscopy images
Specificity
Sensitivity
1
0
0
1
Algorithm: AUC = 0.94

A MATLAB implementation of the TensorFlow Neural Network Playground

version 1.3 (57.6 KB) by Raymond Phan

A MATLAB implementation of the TensorFlow Neural Networks Playground

Tensorflow for Matlab

version 1.1 (472 KB) by Louis Yu

GUI Interface to allow Matlab Users Auto-code to Tensorflow's Python API

https://www.mathworks.com/matlabcentral/fileexchange/62808-tensorflow-for-matlab