

Supplementary information

Ensemble of Self-supervised Learning Methods for Robust Skin Disease Image Diagnosis Leveraging Unlabeled Data

Ryu Tadokoro¹, Kaname Kojima^{2,3}, Kengo Kinoshita^{2,4,5,6}, Yoshihide Asano⁷, Kenshi Yamasaki^{7,8}, Kosuke Shido^{7,9}

1 Tohoku University School of Medicine, 2-1 Seiryomachi, Aoba-ku, Sendai, Miyagi 980-8575, Japan

2 Tohoku Medical Megabank Organization, Tohoku University, 2-1 Seiryomachi, Aoba-ku, Sendai, Miyagi 980-8573, Japan

3 RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

4 Graduate School of Information Sciences, Tohoku University, 6-3-09 Aza-Aoba, Aramaki, Aoba-ku, Sendai, Miyagi 980-8579, Japan

5 Advanced Research Center for Innovations in Next-Generation Medicine, Tohoku University, 2-1 Seiryomachi, Aoba-ku, Sendai, Miyagi 980-0873, Japan

6 Institute of Development, Aging and Cancer, Tohoku University, 4-1 Seiryomachi, Aoba-ku, Sendai, Miyagi 980-8575, Japan

7 Department of Dermatology, Graduate School of Medicine, Tohoku University, 1-1 Seiryomachi, Aoba-ku, Sendai, Miyagi 980-8574, Japan

8 Aloop Clinic & Lab, 1-7-7 Ginza, Chuo-ku, Tokyo 104-0061, Japan

9 Athnomedical, Inc., 2-4-1 Ichiban-cho, Aoba-ku, Sendai, Miyagi 980-0811, Japan.

1 Skin disease image datasets

NSDD: A comprehensive skin disease image dataset from Japanese hospitals

The National Skin Disease Database of Japan (NSDD) is a collection of 390,425 skin disease images from 20,367 patients. These images were sourced from 16 university hospitals across Japan with the support of the Japan Agency for Medical Research and Development (Supplementary Fig. 1a). They were captured using over 85 different devices, including the Canon EOS 10D and NIKON D5000, and primarily feature Japanese patients. A subset of the NSDD, referred to as NSDD-labeled, includes 70,196 images corresponding to 13,038 patients. These images were curated by multiple dermatologists who verified the accuracy of the labels and the presence of lesions, with some images being relabeled as necessary. This subset covers 59 skin diseases (Supplementary Fig. 1b) and was utilized to measure the transfer learning accuracy of the pre-trained models. The remaining images, which lack adequate label curation, were treated as unlabeled data for self-supervised pre-training. Any images overlapping with those in the NSDD-labeled dataset were removed. From these, we randomly selected 100,000 images to create the NSDD-unlabeled dataset. The NSDD-labeled dataset was further divided into three parts: 56,240 images (10,419 patients) for training, 6,918 images (1,303 patients) for validation, and 7,038 images (1,304 patients) for testing.

SD-198: A verified image dataset for skin disease recognition

The SD-198 dataset was used for automatic skin disease recognition and diagnosis. The dataset was collected from DermQuest (<http://www.dermquest.com>), a database of skin disease images. The annotations in this dataset were verified by multiple dermatologists, making it a reliable dataset (DermQuest, 2012). The dataset consists of 6,584 images of 198 diseases, including actinic keratosis, pigmentation, acne vulgaris, and sebaceous gland hyperplasia. The images were taken with digital cameras, smartphones, and other devices, and under various photographing conditions such as lighting and distance from the disease. The dataset targets patients with various conditions such as gender, age, disease site, and skin color. The dataset was divided into training, validation, and testing sets using a 5-fold cross-validation approach. Specifically, 4,040 images were used for training, 1,290 images for validation, and 1,254 images for testing.

Fitzpatrick 17k: A large-scale skin disease image dataset with double-checked annotations

The Fitzpatrick 17k dataset was used for skin disease recognition and diagnosis. The dataset consists of a total of 16,536 images, 12,672 images collected from DermaAmin, and 3,905 images collected from Atlas Dermatologico. The dataset includes 114 major skin diseases, such as psoriasis, squamous cell carcinoma, lichen planus, and basal cell carcinoma. The annotations in this dataset were double-checked by a board-certified dermatologist. The dataset was labeled following the Fitzpatrick classification, with annotations decided by majority vote from 2 - 5 annotators per image. The dataset includes a wide variety of distributions in terms of patient and imaging conditions. The dataset was divided into training, validation, and testing sets using a 5-fold cross-validation approach. Specifically, 9,995 images were used for training, 3,278 images for validation, and 3,263 images for testing.

Derm7pt: a dual image type dataset for skin disease discrimination

The Derm7pt dataset consists of 2,022 skin disease images, including melanoma, and was used for discrimination using a 7-point checklist. The dataset includes two types of images: clinical images and dermoscopic images. The clinical images were taken close to the diseased area, but the imaging conditions were not standardized compared to dermoscopic images. The dermoscopic images were taken under standardized lighting conditions and field of view. The annotations and 7-point checklist criteria attached to each image were annotated by a dermatologist. The dataset was divided into two sets: Derm7pt-clinic and Derm7pt-derm, to separately verify the accuracy changes due to image properties. While the dataset contains images of 14 different diseases, including basal cell carcinoma and lentigo, due the limited number of images, we considered solely a binary classification task, distinguishing melanoma (252 images) and other diseases (759 images). The dataset was divided into training, validation, and testing sets, with 413 images allocated for training, 203 images for validation, and 395 images for testing.

HAM10000: A vast collection of 10,015 dermoscopic images for skin disease analysis and diagnosis

The HAM10000 dataset is a collection of 10,015 dermoscopic images of skin disease, accessible from the ISIC repository (<https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main>). More than half of the disease labels were confirmed by pathology biopsy, while other images were

labeled by either follow-up examination, expert consensus, or confirmation by in-vivo confocal microscopy. The target diseases include actinic keratosis and intraepithelial carcinoma / Bowen’s disease, basal cell carcinoma, benign keratosis-like lesions (solar lentigines / seborrheic keratosis and lichen-planus like keratosis), dermatofibroma, melanoma, melanocytic nevi, and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas, and hemorrhage). The dataset was divided into training, validation, and testing sets, with 10,015 images of publicly available training data used for the division since the official test dataset is not publicly available. Since multiple images taken for one diseased part are tracked as lesion_id, the training set was split into 6,009 images, the validation set to 2,003 images, and the test set to 2,003 images to ensure that lesion_id does not overlap.

Supplementary Table 1: Numbers of training images, validation images, and test images (along with numbers of patients if available) for the evaluation of pre-trained models on each skin disease image dataset.

Dataset	Training images		Validation images		Test images	
	No. of images	No. of patients	No. of images	No. of patients	No. of images	No. of patients
NSDD-labeled	56,240	10,419	6,918	1,303	7,038	1,304
SD-198	4,040	—	1,290	—	1,254	—
Fitzpatrick 17k	9,995	—	3,278	—	3,263	—
Derm7pt-clinic	413	—	203	—	395	—
Derm7pt-derm	413	—	203	—	395	—
HAM1000	6,009	—	2,003	—	2,003	—

Supplementary Table 2: Comparison of skin disease classification accuracies among various pre-trained models fine-tuned with the NSDD-labeled dataset as well as the case of using ensemble approach for these models on external skin disease image datasets.

a SD198

Pre-trained model or method	Image-based classification accuracy (%)		
	Top 1	Top 3	Top 5
Fine-tuned baseline pre-trained model	63.1 (+ 2.8)	78.5 (+ 0.7)	84.8 (+ 0.7)
Hybrid pre-trained model (SimCLR)	63.8 (+ 3.5)	80.0 (+ 2.2)	85.5 (+ 1.4)
Hybrid pre-trained model (BYOL)	64.8 (+ 4.5)	79.9 (+ 2.1)	85.9 (+ 1.8)
Hybrid pre-trained model (DINO)	63.4 (+ 3.1)	79.4 (+ 1.6)	85.1 (+ 1.0)
Ensemble	68.6 (+ 8.3)	83.7 (+ 5.9)	88.3 (+ 4.2)

b Fitzpatrick 17k

Pre-trained model or method	Image-based classification accuracy (%)		
	Top 1	Top 3	Top 5
Fine-tuned baseline pre-trained model	51.8 (+ 0.8)	68.7 (+ 1.1)	75.0 (+ 0.9)
Hybrid pre-trained model (SimCLR)	52.0 (+ 1.0)	68.2 (+ 0.6)	75.5 (+ 1.4)
Hybrid pre-trained model (BYOL)	54.6 (+ 3.6)	70.2 (+ 2.6)	75.7 (+ 1.6)
Hybrid pre-trained model (DINO)	53.3 (+ 2.3)	69.7 (+ 2.1)	75.8 (+ 1.7)
Ensemble	58.7 (+ 7.7)	73.2 (+ 5.6)	79.1 (+ 5.0)

c Derm7pt-clinic

Pre-trained model or method	Image-based classification accuracy (%)
	Top 1
Fine-tuned baseline pre-trained model	75.4 (− 1.1)
Hybrid pre-trained model (SimCLR)	75.9 (− 0.6)
Hybrid pre-trained model (BYOL)	77.0 (+ 0.5)
Hybrid pre-trained model (DINO)	77.0 (+ 0.5)
Ensemble	78.5 (+ 2.0)

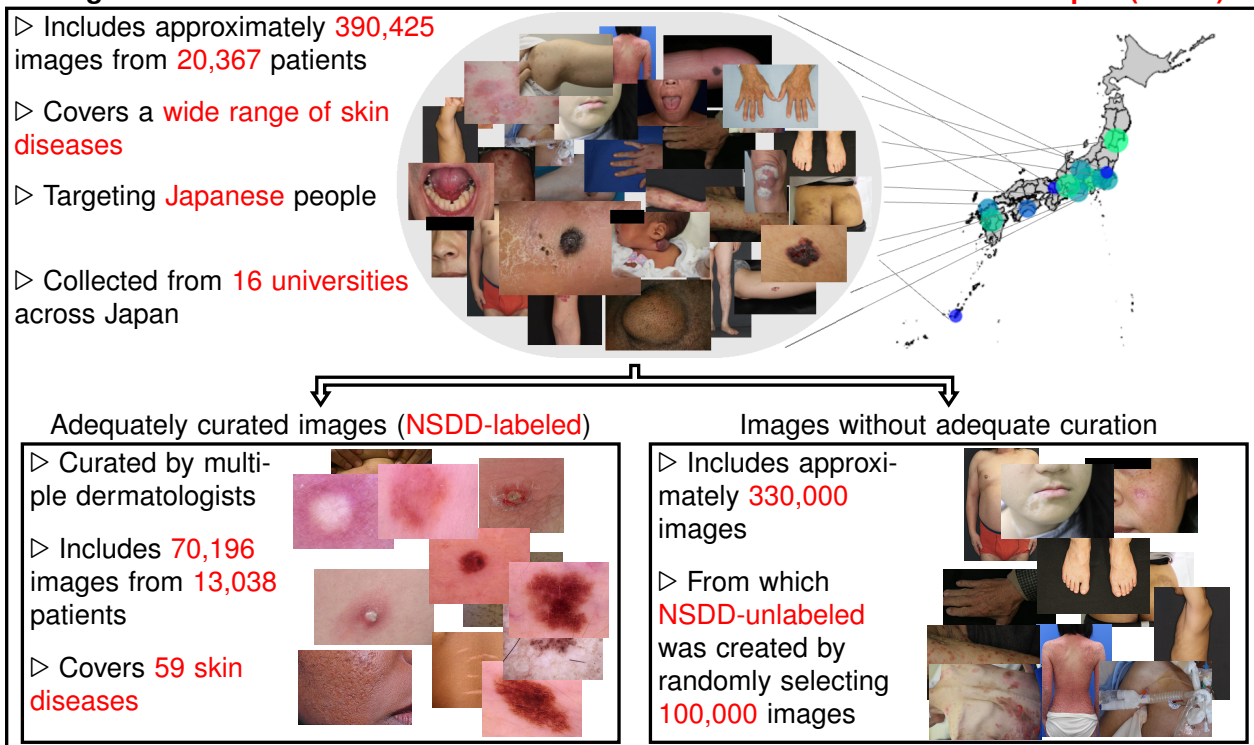
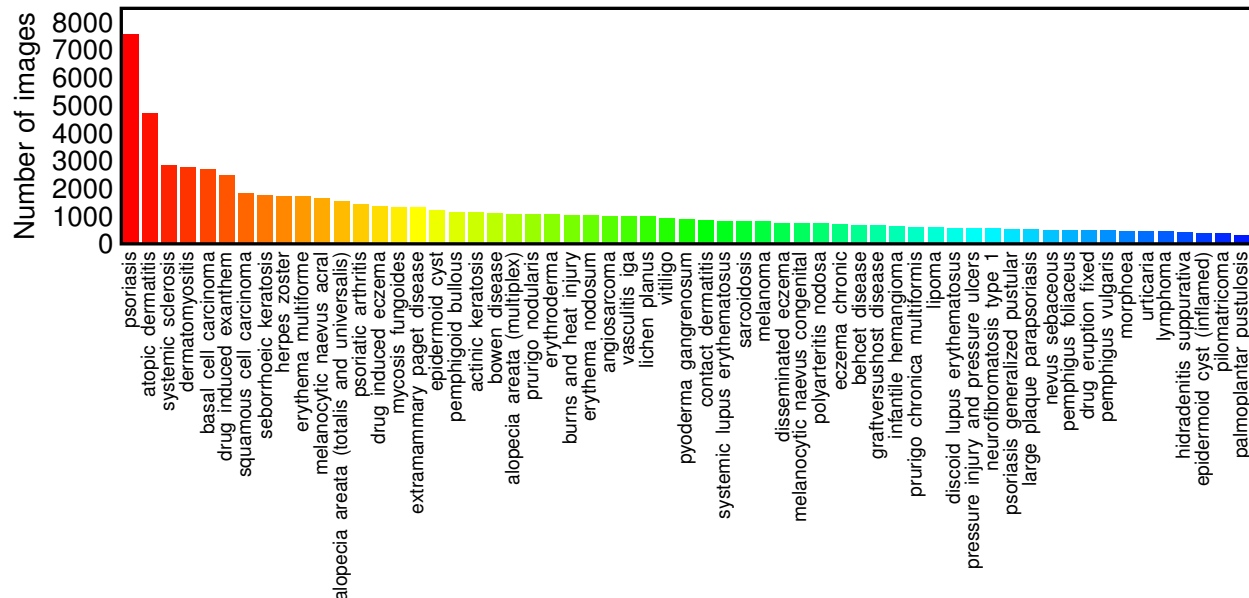
d Derm7pt-derm

Pre-trained model or method	Image-based classification accuracy (%)
	Top 1
Fine-tuned baseline pre-trained model	82.3 (+ 2.3)
Hybrid pre-trained model (SimCLR)	82.3 (+ 2.3)
Hybrid pre-trained model (BYOL)	80.5 (+ 0.5)
Hybrid pre-trained model (DINO)	80.0 (+ 0.0)
Ensemble	82.3 (+ 2.3)

e HAM10000

Pre-trained model or method	Image-based classification accuracy (%)		
	Top 1	Top 3	Top 5
Fine-tuned baseline pre-trained model	88.4 (+ 2.5)	98.5 (+ 0.1)	99.8 (− 0.1)
Hybrid pre-trained model (SimCLR)	89.2 (+ 3.3)	98.5 (+ 0.1)	99.7 (− 0.2)
Hybrid pre-trained model (BYOL)	88.8 (+ 2.9)	98.6 (+ 0.2)	100.0 (+ 0.1)
Hybrid pre-trained model (DINO)	87.8 (+ 1.9)	98.4 (+ 0.0)	99.8 (− 0.1)
Ensemble	89.3 (+ 3.4)	98.5 (+ 0.1)	100.0 (+ 0.1)

Values in parentheses: Indicate the improvement in accuracy from the baseline pre-trained model.

a**Large Scale Dataset of Skin Diseases – National Skin Disease Database of Japan (NSDD)****b****Supplementary Fig. 1: Overview of the National Skin Disease Database of Japan (NSDD).**

a An overview of the NSDD, which contains approximately 400,000 skin disease images. Within the dataset, a subset composed of 70,196 images that correspond to 13,038 patients and cover 59 skin diseases was carefully curated by multiple dermatologists. This subset is referred to as NSDD-labeled. The remaining images, numbering approximately 330,000, are not yet adequately curated, and NSDD-unlabeled was created by randomly selecting 100,000 images from these remaining images.

b A histogram showing the distribution of image counts across the 59 diseases contained within the NSDD-labeled dataset.



Supplementary Fig. 2: Examples of images with rulers in the Derm7pt-clinic dataset.

As images in the Derm7pt-clinic dataset were captured at a close range, rulers used for measuring lesion sizes were often included in the frame.