

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

Kaname Kojima (✉ kojima@megabank.tohoku.ac.jp)

Tohoku Medical Megabank Organization, Tohoku University <https://orcid.org/0000-0001-5713-2483>

Ryu Tadokoro

Tohoku University School of Medicine

Kengo Kinoshita

Department of System Bioinformatics, Graduate School of Information Sciences, Tohoku University

Yoshihide Asano

Graduate School of Medicine, Tohoku University

Kenshi Yamasaki

Graduate School of Medicine, Tohoku University

Kosuke Shido

Tohoku University <https://orcid.org/0000-0002-7764-9196>

Article

Keywords:

Posted Date: September 18th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3288218/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

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.

*To whom correspondence should be addressed; E-mail: kojima@megabank.tohoku.ac.jp and kenshi.yamasaki@gmail.com.

22 Abstract

23 Deep learning technologies have led to remarkable improvements in medical image analysis, but the
 24 collection and annotation of medical data remain challenging. This study leverages self-supervised
 25 learning and unlabeled images from the National Skin Disease Database of Japan (NSDD) to enhance
 26 skin disease classification. By generating pre-trained models using three self-supervised learning meth-
 27 ods, and comparing them to a baseline pre-trained model on ImageNet, we found that pre-training with
 28 unlabeled images increased transfer learning performance by 1-2% on labeled images in the NSDD.
 29 We also demonstrated the effectiveness of the pre-trained models across four public dermatological
 30 image datasets. We further implemented a hybrid pre-training approach that combines self-supervised
 31 and supervised learning, and observed accuracy improvements in certain datasets. Since the optimal
 32 strategy is task and dataset-specific, we proposed an ensemble of transfer learned models from diverse
 33 pre-training methods for robust classification. The ensemble approach consistently enhanced accuracy
 34 across various datasets, with an improvement by up to 8.3% from the baseline, confirming the effec-
 35 tiveness of self-supervised pre-training with unlabeled dermatological images even when the images
 36 are mainly from one ethnic group.

37 Introduction

38 Advancements in deep learning (DL) technology have significantly improved the accuracy of medical
 39 data analysis, including in the field of dermatology. Initially, DL-based methods were developed for bi-
 40 nary classification tasks, such as distinguishing between malignant and benign skin diseases[1, 2, 3, 4].
 41 In tasks such as differentiating malignant melanomas from benign nevi or squamous cell carcinomas
 42 from benign seborrheic keratoses, an existing DL-based method has demonstrated comparable classi-
 43 fication accuracy to that of dermatologists[2]. As the technology has advanced, the scope of DL-based
 44 methods has broadened beyond binary classifications, and current research is focused on developing
 45 methods capable of concurrently distinguishing a wider variety of skin diseases [5, 6, 7]. Our research
 46 group has contributed to this effort by developing a DL-based method that can distinguish 26 types
 47 of skin diseases using the National Skin Disease Database of Japan (NSDD), a large database of
 48 skin disease images collected from 16 Japanese universities, including our institution (Supplementary
 49 Fig. 1a) [8, 9]. Despite these advances, the development of DL-based methods for dermatological
 50 images remains challenging due to the extensive variety of skin diseases; more than 2,000 skin disease

names are recorded in major dermatological textbooks.

In order to handle the extensive variety of skin diseases, especially in diagnosing rare diseases, the collect of large amounts of high-quality image data is essential, but the collection of such data is hampered by several factors. One is the various restrictions on patient data enforced by personal information protection laws. Also, while common digital cameras are widely used in dermatology, obtaining medical image data can present difficulties due to the emergence of new specialized imaging devices and methods, as well as their high imaging costs. Furthermore, annotating images with disease labels is a laborious and time-consuming process. The NSDD was established to address these data-related issues. The NSDD currently contains approximately 400,000 images covering more than 1,500 skin diseases, including rare diseases and provides a comprehensive dataset for developing DL-based methods for various purposes. However, while collecting skin disease data is relatively straightforward, curating and annotating this large dataset is also challenging due to the requirement of substantial human resources. For instance, out of approximately 400,000 images in the NSDD, only 70,196 images have been relabeled and curated by multiple dermatologists. The remaining images, numbering approximately 330,000, are not yet adequately curated.

In this context, self-supervised learning has emerged as a promising solution to the problem of data shortages [10, 11, 12, 13, 14, 15]. This method leverages unlabeled data to enhance the accuracy of DL models. The approach involves representation learning, where unique labels are assigned to each image, and the model is trained to solve ‘pseudo-tasks’ to extract features that represent the characteristics of the data. The model, pre-trained through this process, can then be further fine-tuned for specific tasks, such as diagnosing skin diseases. While the majority of images in the NSDD are not adequately curated, they can be utilized as unlabeled data within this self-supervised learning framework.

In this study, we employ three self-supervised learning methods – SimCLR [11], BYOL [12], and DINO [10] – to generate pre-trained models using 100,000 randomly extracted images from the NSDD that lack adequate label curation. We refer to the dataset comprising these 100,000 images as NSDD-unlabeled. We assess the effectiveness of these pre-trained models in classifying skin diseases through the comparison with a baseline model pre-trained on ImageNet [16]. ImageNet is a widely-used dataset for general object classifications, such as vehicles and animals. This baseline pre-trained model also provides the initial weights for self-supervised learning. For all models in this study, we use ResNet-18 [17] as the underlying architecture. Fig. 1 presents a schematic representation of both

the pre-training process (Fig. 1 a and c) and the evaluation procedure for the pre-trained models considered in this study (Fig. 1 b and d).

We first evaluate the efficacy of our self-supervised pre-trained models on in-distribution data, using 70,196 well-curated images for 59 skin diseases from the NSDD, which we refer to as NSDD-labeled (Supplementary Fig. 1 a and b). We next examine the performance of these models on out-of-distribution data using the following external dermatological image datasets: SD-198 [18], Fitzpatrick 17k [19], Derm7pt [20], and HAM10000 [21] (Fig. 1 b). We further fine-tune these pre-trained models as well as the baseline pre-trained model using the NSDD-labeled dataset (Fig. 1 c) and measure their accuracy on the aforementioned external dermatological image datasets. From these comparisons, we find that the effectiveness of self-supervised learning methods can vary depending on the dataset (Fig. 1 d). Based on these findings, we propose leveraging an ensemble of transfer learned models derived from a variety of pre-training methods, aiming to achieve robust performance, regardless of the target dataset. To substantiate our proposal, we evaluate its performance across these datasets (Fig. 1 d).

Results

Performance comparison of self-supervised learning methods for skin disease classification on the NSDD-labeled dataset

We assessed the efficacy of pre-trained models obtained from self-supervised learning methods in skin disease classification on in-distribution data using the NSDD-labeled dataset. This dataset was partitioned into training, validation, and testing subsets (Supplementary Table 1). The training and validation subsets were used for transfer learning of both self-supervised pre-trained models and the baseline pre-trained model. The transfer learning accuracy of these pre-trained models was then evaluated using the testing subset (Table 1). As multiple images per patient were available, we also calculated patient-based accuracy by averaging prediction values for each patient in addition to image-based accuracy. Accuracy improvements were observed in almost all indicators for the models pre-trained with the NSDD-unlabeled dataset, compared to the baseline pre-trained model. Specifically, the use of the self-supervised pre-trained model with BYOL showed a 1.7% improvement in top-1 image-based accuracy and a 2.2% improvement in top-1 patient-based accuracy compared to the case of using the baseline pre-trained model.

Evaluation of pre-trained models on external dermatological image datasets

We assessed the effectiveness of the self-supervised pre-trained models from the NSDD-unlabeled dataset on out-of-distribution data using the following external dermatological image datasets: SD-198, Fitzpatrick 17k, Derm7pt, and HAM10000 (Table 2). Since patient IDs were not available for images of the above datasets, only image-based accuracies were considered for the evaluation.

Table 2a shows the transfer learning accuracy of each pre-trained model for SD-198. All the self-supervised learning methods showed an improvement in accuracy from the baseline pre-trained model, with varying degrees of improvement depending on the method. DINO had the highest improvement in top-1 accuracy at 3.3%, while SimCLR had the smallest improvement at 0.3%. Table 2b shows the transfer learning accuracy of each pre-trained model for Fitzpatrick 17k. In terms of top-1 accuracy, all the self-supervised learning methods showed an improvement over the baseline. BYOL had the highest improvement at 3.0%, while SimCLR had the smallest improvement at 0.3%. However, some methods had lower accuracy than the baseline pre-trained model in top-3 and top-5 accuracy.

In the Derm7pt dataset, two types of images were taken from each patient: clinical images and dermoscopic images. Clinical images were taken in close proximity to the disease site, but the imaging conditions were less standardized compared to the dermoscopic images, which were captured under consistent lighting conditions and fields of view. In order to separately evaluate the effectiveness of self-supervised pre-trained models on these two types of images, each with distinct characteristics, we divided the Derm7pt dataset into two subsets: Derm7pt-clinic, which is composed of clinical images, and Derm7pt-derm, which is composed of dermoscopic images. Note that we evaluated only the top-1 accuracy for both Derm7pt-clinic and Derm7pt-derm, as these datasets were considered for binary classification tasks only. Table 2c and d show the transfer learning accuracy of each pre-trained model for the Derm7pt-clinic and Derm7pt-derm datasets, respectively. For the Derm7pt-clinic dataset, the self-supervised pre-trained model by BYOL was the most effective, improving accuracy by 1.2%, while the self-supervised pre-trained model by DINO decreased accuracy by 0.6%. On the other hand, for the Derm7pt-derm dataset, the self-supervised pre-trained model by SimCLR was the most effective, with an accuracy improvement of 0.8%, while the self-supervised pre-trained model by BYOL decreased accuracy by 0.2%.

Table 2e shows the transfer learning accuracy of each pre-trained model for HAM10000. All methods showed an improvement in accuracy from the baseline, with varying degrees of improvement depending on the method. BYOL had the highest improvement at 2.9%, while DINO had the second

highest improvement at 1.2%. SimCLR had the smallest improvement at 0.7%. These results suggest that the choice of pre-training method can affect the accuracy improvement in the out-of-distribution data.

Accuracy discrepancies across skin diseases with pre-trained models on the NSDD-labeled dataset

We examined the discrepancies in image-based accuracy between the baseline pre-trained model and each self-supervised pre-trained model on the NSDD-labeled dataset, breaking it down according to individual skin disease labels (Fig. 2). The horizontal axis of the graph in Fig. 2 shows the disease labels, ordered by the volume of available data in descending order, while the vertical axis indicates discrepancies in image-based accuracy. There was a noticeable trend indicating that the accuracy for certain diseases is likely to increase or decrease, irrespective of the self-supervised learning method used. Disease labels that consistently demonstrated an improvement in accuracy across all three self-supervised learning methods included drug-induced exanthem (including both morbilliform drug eruption and maculopapular drug eruption), squamous cell carcinoma, Bowen’s disease, vitiligo, erythema nodosum, infantile hemangioma, neurofibromatosis type 1, lipoma, discoid lupus erythematosus, sebaceous nevus, pressure injury, and pressure ulcers (blue bars and dotted lines in Fig. 2). On the contrary, alopecia areata (totalis and universalis) displayed a consistent reduction in accuracy across all three self-supervised learning methods (a red bar and a dotted line in Fig. 2). Similarly, drug-induced eczema, actinic keratosis, and epidermoid cyst (inflamed) also exhibited reduced accuracy with all three self-supervised learning methods (red bars and dotted lines in Fig. 2).

Performance verification of additional pre-training with the NSDD-labeled dataset on external skin disease image datasets

We prepared hybrid pre-trained models by fine-tuning self-supervised pre-trained models using the NSDD-labeled dataset in order to examine the effectiveness of additional pre-training with the NSDD-labeled dataset. We then evaluated their effectiveness on the SD-198, Fitzpatrick 17k, Derm7pt-clinic, Derm7pt-derm, and HAM10000 datasets (Fig. 3a, Supplementary Table 2). For comparison, we also considered the fine-tuned baseline pre-trained model with the NSDD-labeled dataset. When using the fine-tuned baseline pre-trained model as a pre-trained model for transfer learning, we observed an accuracy improvement of 2.8% for the SD-198 dataset and 0.8% for Fitzpatrick 17k, compared to the

baseline pre-trained model. The hybrid pre-trained models further increased accuracy by up to 1.7% in SD-198 and up to 2.8% in Fitzpatrick 17k.

In contrast, for the Derm7pt-clinic dataset, using the fine-tuned baseline pre-trained model resulted in a 1.1% decrease in accuracy compared to the baseline pre-trained model. When using hybrid pre-trained models, a reduction in accuracy of 0.6% was observed in the case of using SimCLR as the self-supervised learning method, while accuracy improvements of 0.5% were observed for the cases using BYOL and DINO. One potential reason for the decreased accuracies in the fine-tuned baseline pre-trained model and the hybrid pre-trained model using SimCLR could be the specific characteristics of the Derm7pt-clinic dataset, which is composed of close-up clinical images. Some of these images include rulers, as shown in Supplementary Fig. 2, that are used for measuring lesion sizes. Since images in the NSDD-unlabeled dataset were not adequately curated, unnecessary objects like rulers were included in some images. Consequently, further fine-tuning with well-curated images, such as those in the NSDD-labeled dataset, may diminish the robustness of the model when it faces images containing such extraneous elements.

In the Derm7pt-derm dataset, the fine-tuned baseline pre-trained model improved accuracy by 2.3% compared to the baseline pre-trained model. When using SimCLR and DINO as self-supervised learning methods in hybrid pre-training, improvements in accuracy by 2.3% and 0.5% were observed, respectively. On the other hand, no accuracy improvement was observed when using BYOL as a self-supervised learning method. In the case of the HAM10000 dataset, the fine-tuned baseline pre-trained model led to a general improvement in accuracy, and a remarkable improvement of 2.7% was observed for the hybrid pre-trained model when using SimCLR as a self-supervised learning method.

For datasets with varying shooting conditions, such as SD-198 and Fitzpatrick 17k, the hybrid pre-trained models improved the accuracy of all self-supervised learning methods compared to the fine-tuned baseline pre-trained model. However, for dermoscopic image datasets like the Derm7pt-derm and HAM10000 datasets, while the fine-tuned baseline pre-trained model demonstrated higher accuracy than the baseline pre-trained model, the effectiveness of the hybrid pre-trained models depended on the specific self-supervised learning method used.

Performance verification of ensemble of transfer learned models on external skin disease image datasets

Performance evaluations have thus far shown that the efficacy of self-supervised learning methods can vary depending on the dataset and target disease type. With this in mind, we proposed an ensemble of models, each obtained from transfer learning of various hybrid pre-trained models, with the aim of achieving more robust classification. We then validated the effectiveness of this approach on each of the aforementioned external datasets (Fig. 3b, Supplementary Table 2). The ensemble approach improved the accuracy over the baseline pre-trained model by 8.3%, 7.7%, 2.0%, 2.3%, and 3.4% for the SD-198, Fitzpatrick 17k, Derm7pt-clinic, Derm7pt-derm, and HAM10000 datasets, respectively. Interestingly, the ensemble approach provided better accuracy than the best accuracy observed in the cases of both hybrid pre-trained models and the fine-tuned baseline pre-trained model for the datasets other than the Derm7pt-derm dataset. The improvements over the best accuracies were 3.8%, 4.1%, 1.5%, and 0.1% for the SD-198, Fitzpatrick 17k, Derm7pt-clinic, HAM10000 datasets, respectively. Even for the Derm7pt-derm dataset, the ensemble approach provided accuracy as high as the best one among the cases of hybrid pre-trained models and the fine-tuned baseline pre-trained model. These results demonstrate a consistent performance through the ensemble approach and its ability to achieve robust skin disease classification across diverse datasets.

Discussion

This study assessed the effectiveness of self-supervised learning methods for pre-training in medical image analysis tasks. Our results suggest that pre-training with unlabeled datasets, particularly those with similar properties to the target datasets, can enhance performance in transfer learning. For instance, in the NSDD-labeled dataset, self-supervised learning methods exhibited about 1 to 2% higher accuracy than the baseline pre-trained model, which was obtained from supervised learning with ImageNet. Patient-based accuracy, a metric aggregating predictions from multiple images of the same patient to make the final diagnosis, peaked at 60.1%, consistently outperforming image-based accuracy which peaked at 51.0%. This aggregation in patient-based accuracy can be seen as a process similar to test time augmentation [22, 23], where predictions from augmented images are combined for the final decision, thereby improving overall accuracy.

The study revealed that the optimal pre-training method might vary by dataset. For example,

DINO outperformed BYOL on the SD-198 dataset, whereas BYOL excelled over DINO on the Fitzpatrick 17k dataset. These results align with previous study highlighting the improved performance of self-supervised learning methods over pre-trained models [24]. Notably, self-supervised pre-training with the NSDD-unlabeled dataset proved beneficial for diverse datasets, including those containing only close-up and dermoscopic images. Even with differing target ethnic groups between the pre-training and transfer learning datasets, significant accuracy improvement was achievable. This underscores the potential of self-supervised learning for pre-training in medical image analysis tasks, particularly when using similar datasets for pre-training and transfer learning. However, our results also suggest that the training dataset can influence the accuracy change per label, independent of the pre-training method. This highlights the need for careful consideration of potential biases in the pre-training dataset when applying transfer learning [25].

We examined how the decision-making process varied among different pre-trained models by analyzing the regions focused on during decision-making, using Grad-CAM [26]. Fig. 4a and b visualize the original test skin lesion images from the NSDD-labeled dataset, highlighted by Grad-CAM. These figures contrast the results of the baseline pre-trained model with those of the self-supervised pre-trained model using BYOL. We chose BYOL for this Grad-CAM analysis because it achieved the highest accuracy among the three self-supervised learning methods in the NSDD-labeled dataset. In Fig. 4a, skin lesion images of vitiligo and sebaceous conditions are shown, and disease labels were correctly classified in both cases. For vitiligo, skin lesions located on both sides of the navel were correctly highlighted in both cases. A small white spot located in the upper left part of the figure was highlighted more strongly in the case of the self-supervised pre-trained model than with the baseline pre-trained model. For sebaceous conditions, the skin lesion area was correctly highlighted in both cases. While the skin lesion was clearly highlighted when using the self-supervised pre-trained model, hair and medical tape around the skin lesion were also highlighted with the baseline pre-trained model. Fig. 4b displays skin lesion images of lipoma and squamous cell carcinoma, with both cases resulting in misclassified disease labels. For lipoma, the baseline pre-trained model failed to highlight the skin lesion area, while the self-supervised pre-trained model correctly highlighted the area. As the self-supervised model was able to capture the skin lesion area, we anticipate that using more images for fine-tuning could be more beneficial for this model than for the baseline. This discussion aligns with the premise of our study that effective pre-training can enhance accuracy. For squamous cell carcinoma, the marking of the lesion for surgery complicates the classification of its disease label. Despite the incorrect disease

label classification, the self-supervised pre-trained model was still able to correctly highlight the lesion area under such conditions, unlike the baseline pre-trained model. This could be attributed to the inclusion of images showing a skin lesion with marking in the NSDD-unlabeled dataset.

We also explored a hybrid pre-training approach, combining both self-supervised and supervised learning methods. This strategy significantly boosted accuracy for the SD-198 and Fitzpatrick 17k datasets. Conversely, for the Derm7pt and HAM10000 datasets, the self-supervised pre-training with the NSDD-unlabeled dataset provided higher accuracy than the baseline pre-trained model. However, additional fine-tuning with NSDD-labeled did not consistently improve accuracy, implying that the optimal strategy might vary depending on the specific task and dataset. We further considered an ensemble of models, transfer learned from the hybrid pre-trained models for each external dataset. Our results demonstrated that the ensemble approach substantially and consistently enhanced the highest accuracy, with an improvement of up to 8.3% from the baseline pre-trained model, and displayed robustness across varying types of datasets.

In medical data analysis, the accuracy of machine learning methods is vital, while the availability of medical data is often constrained by privacy concerns. This constraint makes it challenging to improve classification accuracy using only publicly available data. While sharing raw medical data may be fraught with legal and ethical concerns, statistical data such as model parameters can be shared publicly. Our self-supervised and hybrid pre-trained models, accessible at <https://github.com/athnomedical/NSDD-SSL>, are designed to enhance classification accuracy without needing raw medical data. Furthermore, our proposed learning strategy serves as a general framework, and we anticipate that it will contribute to various medical studies. While our approach improves overall accuracy across different dermatological image datasets, the importance of each label can vary greatly in the medical domain. For instance, diseases like malignant tumors demand higher precision because of their potential severity, while non-life-threatening conditions such as atopy might not. This disparity emphasizes the need to consider detailed accuracy for each disease instead of focusing solely on overall accuracy, reflecting the varying importance of different medical conditions. Addressing this issue will be an important focus of our future work.

In conclusion, our study has shown the effectiveness of using the National Skin Disease Database (NSDD), a large-scale skin disease image database consisting solely of Japanese patients, for pre-training DL models. While pre-training with unlabeled dermatological images led to an overall improvement in accuracy compared to the baseline pre-trained model, we noticed a decline in accuracy

for certain disease labels, likely due to unexpected biases in the unlabeled image dataset used for pre-training. We addressed this issue by implementing an ensemble approach for robust classification, which mitigated these decreases and consistently outperformed the baseline model across all datasets. Our study hence highlight the potential of unlabeled, large-scale skin image data for pre-training with the use of self-supervised learning methods for analyzing various types of skin disease image datasets, even when that data is mainly from a single ethnic group, as is the case with the NSDD.

Methods

Ethical statements

This research was approved by the ethical committee of the Tohoku University School of Medicine (Reference number: 2021-1-1142). It complies with all relevant ethical regulations.

Skin disease image datasets

We employed a variety of datasets to train and test our models. We used the NSDD-unlabeled dataset, composed of 10,000 images randomly selected from those without adequate label curation in the NSDD, as unlabeled training data to generate self-supervised pre-trained models. The NSDD-labeled dataset, composed of 70,196 images with adequately curated labels, was used to evaluate the effectiveness of the pre-trained models on in-distribution data. For the evaluation on out-of-distribution data, we used the following external dermatological image datasets, which differ from the NSDD-unlabeled datasets in terms of ethnic groups and imaging conditions: the SD-198, Fitzpatrick 17k, Derm7pt, and HAM10000. In the Derm7pt dataset, each patient is represented by two types of images: clinical and dermoscopic. To account for this, we divided the Derm7pt dataset into two subsets, Derm7pt-clinic (comprising the clinical images) and Derm7pt-derm (comprising the dermoscopic images), allowing us to evaluate images with different characteristics separately. The NSDD-labeled, SD-198, Fitzpatrick 17k, Derm7pt-clinic, Derm7pt-derm, and HAM10000 datasets were each partitioned into training, validation, and test images for the evaluation of pre-trained models. Details on the number of images in each set are provided in Supplementary Table 1. More comprehensive information including data preparation about the NSDD, SD-198, Fitzpatrick 17k, Derm7pt, and HAM10000 datasets is in Section 1 of the Supplementary Information.

Implementation details for self-supervised pre-training, fine-tuning, transfer learning, and ensemble learning

We implemented self-supervised pre-training, fine-tuning, transfer learning, and ensemble learning for DL models using Python3 and PyTorch as the deep learning framework. All computations were conducted on a GeForce RTX 3090 GPU. The details of these learning processes are as follows.

Self-supervised pre-training using NSDD-unlabeled We prepared self-supervised pre-trained models using the NSDD-unlabeled dataset, employing ResNet-18 as the architecture of DL models. As state-of-the-art self-supervised learning methods, we employed SimCLR, BYOL, and DINO. All images in the NSDD-unlabeled dataset were used as training data, and the model parameters were updated to minimize the loss function set by each method over a specified number of epochs. For SimCLR, we used LARS [27] as the optimizer and trained the DL model for 100 epochs. The input size was set to 224×224 pixels, with a batch size of 512 and a learning rate of 0.1. Various data augmentations, such as random resized crop, random horizontal flip, color jitter, random grayscale, and normalization, were applied to minimize the contrastive loss. For both BYOL and DINO, we used AdamW [28] as the optimizer and trained the DL model for 100 epochs with an input size of 224×224 pixels, a batch size of 64, and a learning rate of 1.0×10^{-5} . In the case of BYOL, we applied similar data augmentations as with SimCLR to minimize the mean square error between the embeddings of the two images generated by the transformation. For DINO, the same set of augmentations was applied to minimize the cross-entropy loss between the embeddings of the two images generated by the transformation.

Fine-tuning and transfer learning of pre-trained models The same learning procedures were applied in both fine-tuning of hybrid pre-trained models and the baseline pre-trained model, as well as in transfer learning for the evaluation on each dataset. In both cases, we modified the pre-trained models by replacing the linear layer following the global average pooling layer with a new linear layer. This new layer was initialized with weights and designed to have an output size that matched the number of skin disease labels in the target dataset. The parameters of the pre-trained models were then updated by minimizing the cross-entropy loss on the training images from the target dataset, using the Adam optimizer [29], with a batch size of 32 and a learning rate of 1.0×10^{-5} . Training was stopped if the minimum loss for the validation images from the target dataset did not improve after

20 attempts. The model parameters from the epoch with the lowest validation loss were then selected for testing. For data augmentation, we followed the procedures suggested in [30], applying geometric and color transformations to the training images to enhance the accuracy of skin disease classification. We resized the shorter side of the input image to 224 pixels for the SD-198 and Fitzpatrick 17k datasets, and to 448 pixels for other datasets. We then applied horizontal and vertical flipping, random affine transformation, and random luminance change to the resized images as the geometric and color transformations. After the transformation, we cropped it randomly to a size of 224×224 pixels for the SD-198 and Fitzpatrick 17k datasets, or 448×448 pixels for other datasets and normalized the pixel values using the mean and standard deviation of the pixels in the dataset.

Ensemble of transfer learned models derived from diverse pre-training methods. Ensemble of transfer learned models derived from diverse pre-training methods. To ensure more robust classification across the previously mentioned external datasets, we established an ensemble of transfer learned models. This ensemble was achieved by taking the weighted sum of logits from the transfer learned models and feeding this into the softmax function to derive the probability distribution for the classification results. To favor the sum of the weights being one and individual weights being non-negative, we incorporated the following regularization loss in addition to the cross-entropy loss:

$$\sum_i \left(\min(w_i, 0) \right)^2 + \left(\sum_i w_i - 1 \right)^2,$$

where w_i is the weight for the logit of i th model. The weights were initialized to the reciprocal of the total number of the transfer learned models, and then trained similarly to the process employed in transfer learning, using the same training and validation images for each external dataset. Note that the parameters of the transfer learned models were unchanged during this training process.

Data availability

Self-supervised pre-trained models, obtained using the NSDD-unlabeled dataset, and hybrid pre-trained models, using both the NSDD-unlabeled and NSDD-labeled datasets, are available at <https://github.com/athnomedical/NSDD-SSL>.

References

- [1] Chang, W., Huang, A., Yang, C., Lee, C., Chen, Y., Wu, T. & Chen, G. Computer-aided diagnosis of skin lesions using conventional digital photography: a reliability and feasibility study. *PLOS ONE* **8**, e76212 (2013).
- [2] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. & Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
- [3] Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I. & Chang, S.E. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol* **138**, 1529–1538 (2018).
- [4] Upadhyay, P. & Chandra, S. Construction of adaptive pulse coupled neural network for abnormality detection in medical images. *Applied Artificial Intelligence* **32**, 1–19 (2018).
- [5] Lin, J., Guo, Z., Li, D., Hu, X. & Zhang, Y. Automatic classification of clinical skin disease images with additional high-level position information. *Proceedings of the 2019 Chinese Control Conference* 8606–8610 (2019).
- [6] Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G.S., Peng, L.H., Webster, D.R., Ai, D., Huang, S.J., Liu, Y., Dunn, R.C. & Coz, D. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine* **26**, 900–908 (2020).
- [7] Xie, B., He, X., Zhao, S., Li, Y., Su, Z. Zhao, X., Kuang, Y., Wang, Y. & Chen, X. XiangyaDerm: a clinical image dataset of Asian race for skin disease aided diagnosis. *Lecture Notes in Computer Science* **11851**, 22–31 (2019).
- [8] Fujimoto, A., Iwai, Y., Ishikawa, T., Shinkuma, S., Shido, K., Yamasaki, K., Fujisawa, Y., Fujimoto, M., Muramatsu, S. & Abe, R. Deep neural network for early image diagnosis of Stevens-Johnson syndrome/toxic epidermal necrolysis. *J Allergy Clin Immunol Pract* **10**(1), 277–283 (2022).

- [9] Tanaka, M., Saito, A., Shido, K., Fujisawa, Y., Yamasaki, K., Fujimoto, M., Murao, K., Ninomiya, Y., Satoh, S. & Shimizu, A. Classification of large-scale image database of various skin diseases using deep learning. *Int J Comput Assist Radiol Surg* **16**, 1875–1887 (2021).
- [10] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P. & Joulin, A. Emerging properties in self-supervised vision transformers. *Proceedings of the 18th International Conference on Computer Vision* 9630–9640 (2021).
- [11] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning* **149**, 1597–1607, (2020).
- [12] Grill, J-B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B. Kavukcuoglu, K., Munos, R. & Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. *Proceedings of the 34th International Conference on Neural Information Processing Systems* **1786**, 21271–21284 (2020).
- [13] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9726–9735 (2020).
- [14] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpankaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P. & Ng, A.Y. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence* **33**(01), 590–597 (2019).
- [15] Sowrirajan, H., Yang, J., Ng, A. & Rajpurkar, P. MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models. *Proceedings of Machine Learning Research (MIDL Full paper track)* **143**, 727–743 (2021).
- [16] Deng, J., Dong, W., Socher, R., Li, L., Li, K. & Fei-Fei, L. ImageNet: A large-scale hierarchical image database. *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255, (2009).

- [17] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition, *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* 770–778, (2016).
- [18] Sun, X., Yang, J., Sun, M. & Wang, K. A benchmark for automatic visual classification of clinical skin disease images. *Proceeding of the 14th European Conference on Computer Vision*, 206–222, (2016).
- [19] Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A. & Badri, O. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. *Proceeding of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 1820–1828, (2021).
- [20] Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics* **23**, 538–546 (2019).
- [21] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5**, 180161 (2018).
- [22] Kimura, M. Understanding test-time augmentation. *Proceedings of the 28th International Conference on Neural Information Processing* 558–569 (2021).
- [23] Shanmugam, D., Blalock, D.W., Balakrishnan, G. & Gutttag, J.V. Better aggregation in test-time augmentation. *Proceedings of the 18th International Conference on Computer Vision* 1194–1203 (2021).
- [24] Newell, A. & Deng, J. How useful is self-supervised pretraining for visual tasks? *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7343–7352 (2020).
- [25] Salman, H., Jain, S., Ilyas, A., Engstrom, L., Wong, E. & Madry, A. When does bias transfer in transfer learning? *arXiv preprint* arXiv:2207.02842 (2022).
- [26] Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D. & Batra, D.. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* **128**, 336–359 (2017).
- [27] You, Y., Igor Gitman, I. & Ginsburg, B. Large batch training of convolutional networks, *arXiv preprint* arXiv:1708.03888 (2017).

- [28] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *Proceedings of the 7th International Conference on Learning Representations* (2019).
- [29] Kingma, D.P. & Ba, J. Adam: A method for stochastic optimization, *Proceedings of the 3rd International Conference on Learning Representations* (2015).
- [30] Perez, F., Vasconcelos, C., Avila, S. & Valle, E. Data augmentation for skin lesion analysis. *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis (eds Stoyanov D, et al.)* 303–311, Springer (2018).

Acknowledgements

This work was supported by Japan Agency for Medical Research and Development (AMED), Boehringer Ingelheim International GmbH, and Maruho Co., Ltd.

Use of large language models

We used ChatGPT (<https://chat.openai.com>) only for improving the readability of texts.

Author information

Contributions

R.T., K.Ko., K.Y., and K.S. conceptualized and designed the study. R.T. conducted data analysis. R.T., K.Ko., K.Ki, K.Y, and K.S. wrote the manuscript. K.Ki., Y.A, K.Y., and K.S. supervised the study.

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary information

Detailed description about the skin disease image datasets used in this study, supplementary tables, and supplementary figures, are in a PDF file for Supplementary Information.

Table 1: Comparison of skin disease classification accuracies among various pre-trained models on the NSDD-labeled dataset.

Dataset	Pre-training method		Image-based classification accuracy (%)			Patient-based classification accuracy (%)		
			Top 1	Top 3	Top 5	Top 1	Top 3	Top 5
ImageNet	Supervised		49.3	70.3	78.3	57.9	79.8	86.7
NSDD-unlabeled	Self-supervised	SimCLR	50.1 (+ 0.8)	71.3 (+ 1.0)	79.3 (+ 1.0)	59.1 (+ 1.2)	80.0 (+ 0.2)	86.4 (− 0.3)
		BYOL	51.0 (+ 1.7)	71.9 (+ 1.6)	80.1 (+ 1.8)	60.1 (+ 2.2)	80.8 (+ 1.0)	87.5 (+ 0.8)
		DINO	50.2 (+ 0.9)	71.8 (+ 1.5)	79.5 (+ 1.2)	59.4 (+ 1.5)	79.8 (+ 0.0)	87.2 (+ 0.5)

Dataset: Refers to the dataset used for pre-training. Values in parentheses: Indicate the improvement in accuracy compared to the supervised pre-training with ImageNet. Bold values: Highlight the highest accuracy achieved in each category.

Table 2. Comparison of skin disease classification accuracies among various pre-trained models on external skin disease image datasets.

a SD-198

Dataset	Pre-training method		Image-based classification accuracy (%)		
			Top 1	Top 3	Top 5
ImageNet	Supervised		60.3	77.8	84.1
NSDD-unlabeled	Self-supervised	SimCLR	60.6 (+ 0.3)	78.7 (+ 0.9)	84.9 (+ 0.8)
		BYOL	62.1 (+ 1.8)	78.1 (+ 0.3)	84.8 (+ 0.7)
		DINO	63.6 (+ 3.3)	79.1 (+ 1.3)	84.4 (+ 0.3)

b Fitzpatrick 17k

Dataset	Pre-training method		Image-based classification accuracy (%)		
			Top 1	Top 3	Top 5
ImageNet	Supervised		51.0	67.6	74.1
NSDD-unlabeled	Self-supervised	SimCLR	51.3 (+ 0.3)	67.5 (− 0.1)	74.4 (+ 0.3)
		BYOL	54.0 (+ 3.0)	69.4 (+ 1.8)	75.3 (+ 1.2)
		DINO	53.1 (+ 2.1)	67.7 (+ 0.1)	73.8 (− 0.3)

c Derm7pt-clinic

Dataset	Pre-training method	Image-based classification accuracy (%)		
		Top 1		
ImageNet	Supervised	76.5		
NSDD-unlabeled	Self-supervised	SimCLR	77.5 (+ 1.0)	
		BYOL	77.7 (+ 1.2)	
		DINO	75.9 (− 0.6)	

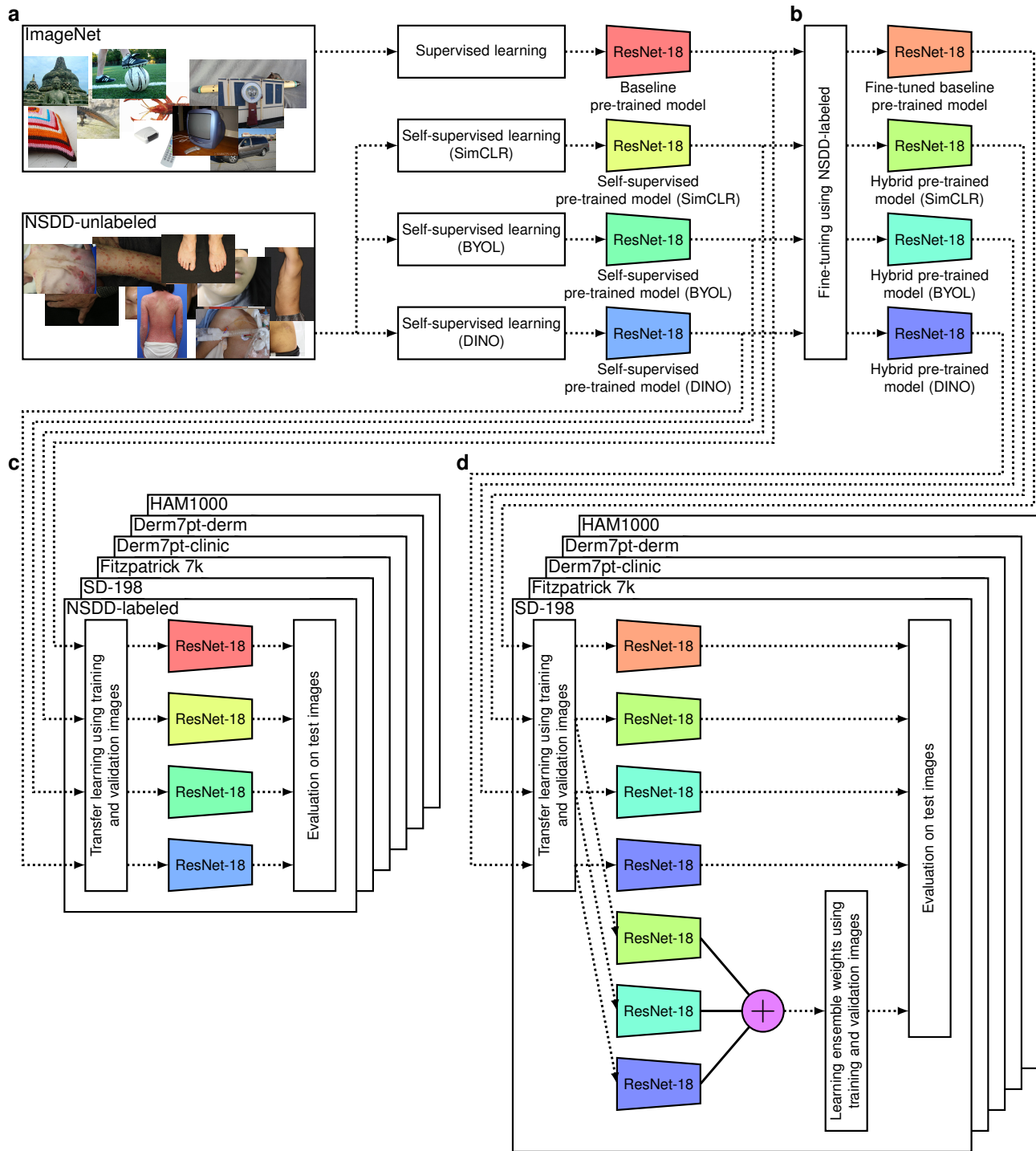
d Derm7pt-derm

Dataset	Pre-training method	Image-based classification accuracy (%)		
		Top 1		
ImageNet	Supervised	80.0		
NSDD-unlabeled	Self-supervised	SimCLR	80.8 (+ 0.8)	
		BYOL	79.8 (− 0.2)	
		DINO	80.3 (+ 0.3)	

e HAM10000

Dataset	Pre-training method		Image-based classification accuracy (%)		
			Top 1	Top 3	Top 5
ImageNet	Supervised		85.9	98.4	99.9
NSDD-unlabeled	Self-supervised	SimCLR	86.6 (+ 0.7)	98.8 (+ 0.4)	100.0 (+ 0.1)
		BYOL	88.8 (+ 2.9)	99.0 (+ 0.6)	100.0 (+ 0.1)
		DINO	87.1 (+ 1.2)	98.6 (+ 0.2)	99.9 (+ 0.0)

497 Values in parentheses: Indicate the improvement in accuracy compared to the supervised pre-training
498 with ImageNet. Bold values: Highlight the highest accuracy achieved in each category.



499

500 **Fig. 1: The schematic diagram illustrating the process of pre-training and evaluation of**
 501 **pre-trained models.**

502 **a** Preparation of the baseline pre-trained model, which is obtained form the supervised learning using
 503 ImageNet, as well as the self-supervised pre-trained models, which uses the NSDD-unlabeled dataset
 504 as unlabeled data.

505 **b** Fine-tuning of both the baseline pre-trained model and self-supervised pre-trained models using the
506 NSDD-labeled dataset to produce the fine-tuned baseline model and hybrid pre-trained models.

507 **c** Evaluation of the pre-trained models from **a**, using an in-distribution dataset (the NSDD-labeled)
508 as well as out-of distribution datasets (SD-198, Fitzpatrick 17k, Derm7pt-clinic, Derm7pt-derm, and
509 HAM1000). Each model (ResNet-18), fine-tuned with training and validation images, is derived from
510 the pre-trained model (ResNet-18) represented by the same color in **a**.

511 **d** Evaluation of the pre-trained models from **b** as well as an ensemble of fine-tuned models, using the
512 out-of distribution datasets. Each model (ResNet-18), fine-tuned with training and validation images,
513 is derived from the pre-trained model (ResNet-18) represented by the same color in **b**.

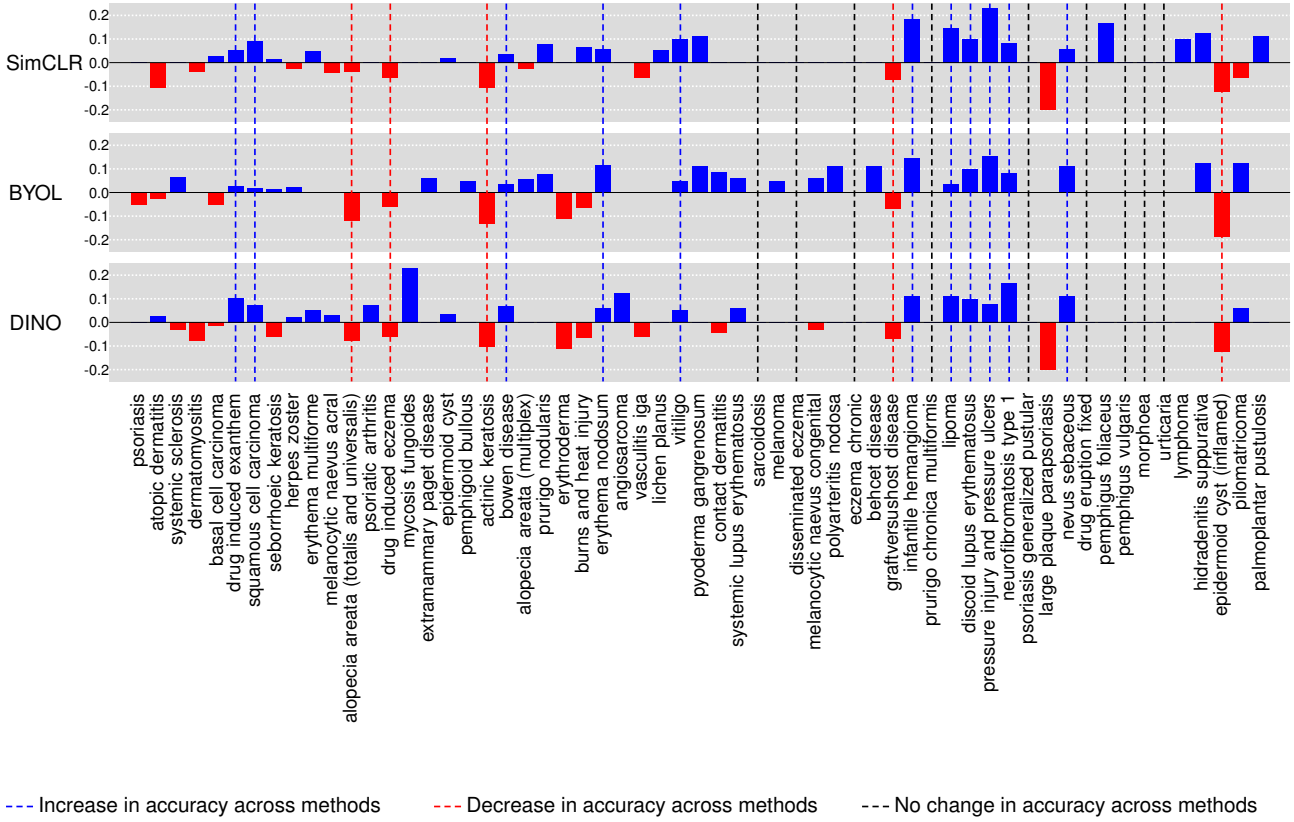


Fig. 2: Changes in image-based classification accuracies due to self-supervised pre-training from the baseline for each skin disease category.

The top, middle, and bottom rows represent the accuracies for self-supervised pre-training by SimCLR, BYOL, and DINO, respectively. The accuracy for each skin disease classification for the case of using the baseline pre-trained model is set as the zero baseline. Increases and decreases in accuracy resulting from self-supervised pre-training by SimCLR, BYOL, and DINO with the NSDD-unlabeled dataset are represented as blue and red bars, respectively. Skin disease names are arranged on the horizontal axis from left to right, in descending order based on the number of images for each disease in the NSDD-labeled dataset. Blue and red dashed lines highlight disease categories where all pre-training methods show consistent trends in accuracy change (increase or decrease, respectively). Black dashed lines denote disease categories without significant accuracy changes across all pre-trained models.

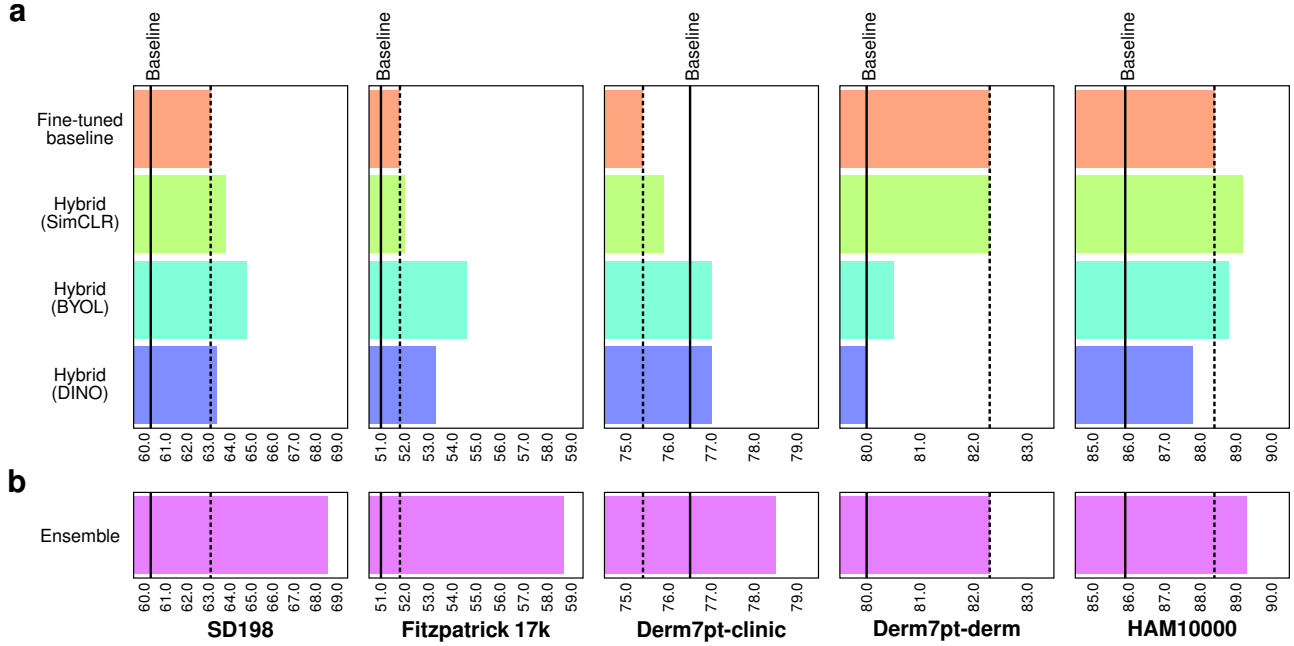
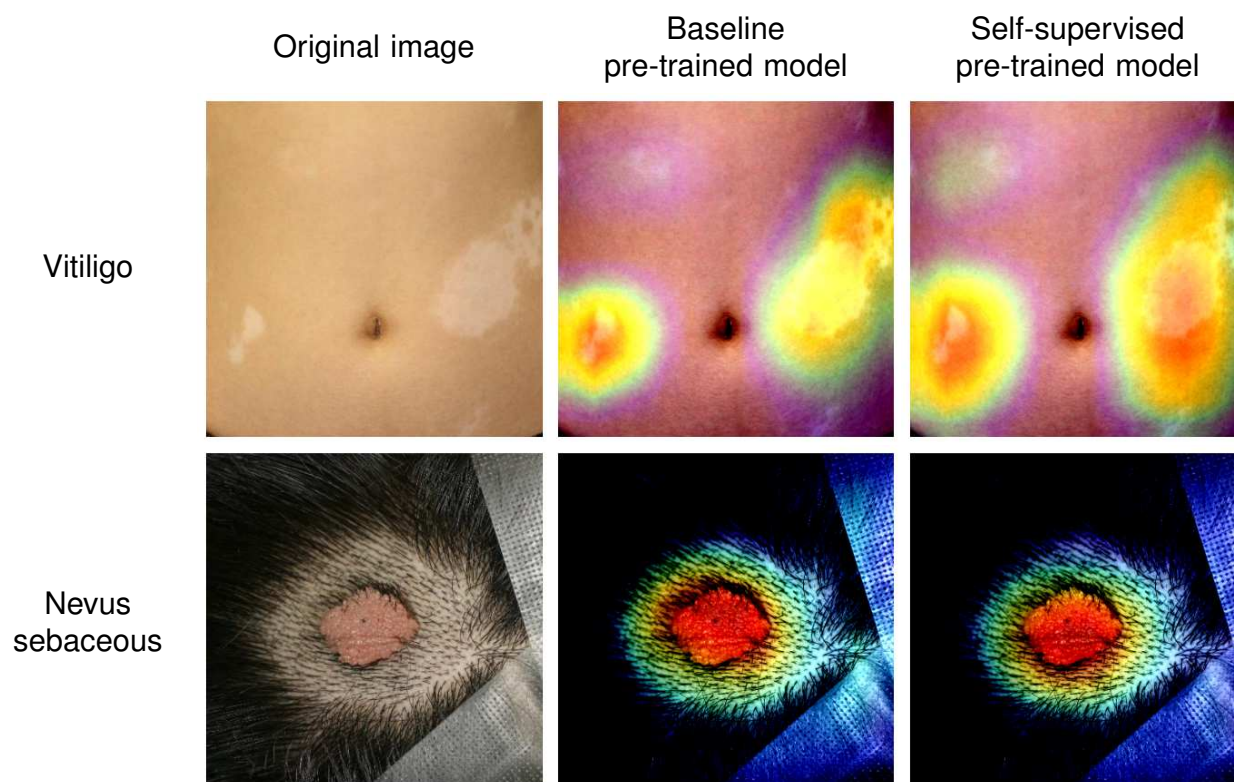
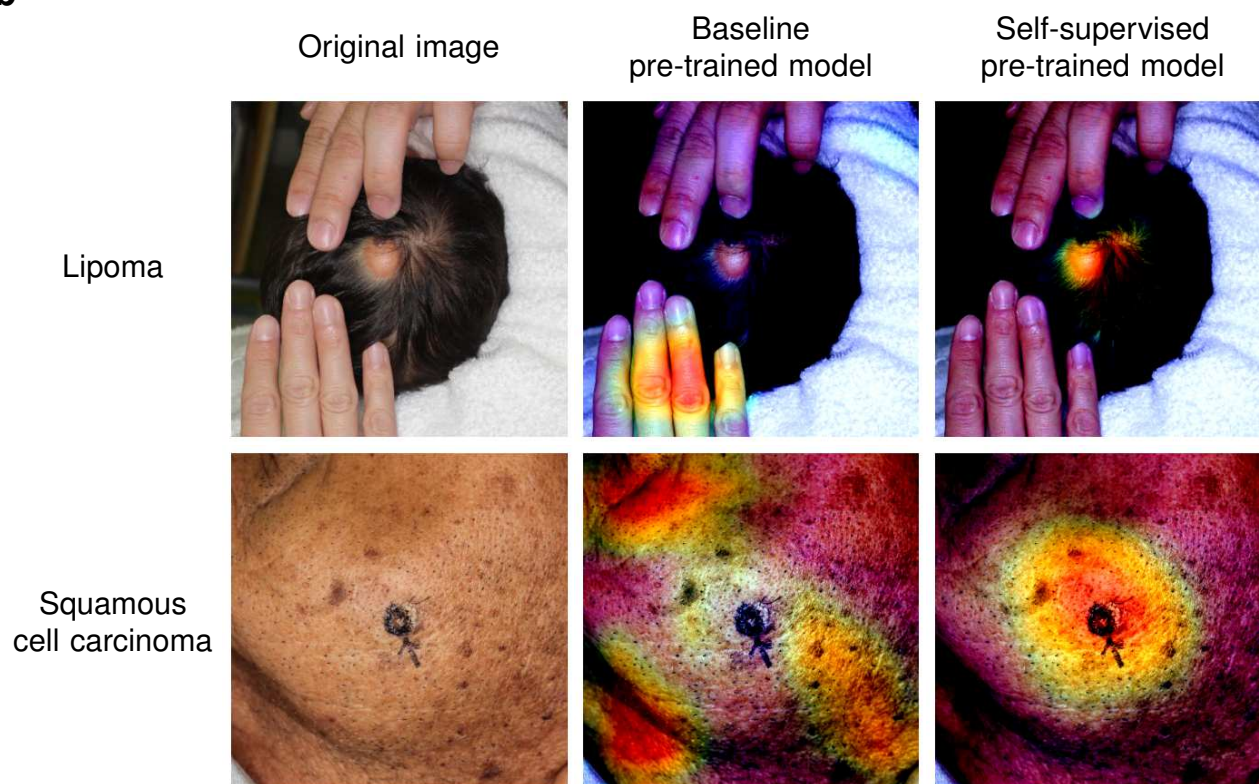


Fig. 3: Comparison of image-based accuracy using pre-trained models fine-tuned with the NSDD-labeled dataset and ensemble approach across out-of-distribution datasets.

a The barplots compare accuracies across out-of-distribution datasets when using the fine-tuned baseline pre-trained model and the hybrid pre-trained models as initial weights. The solid line in each barplot denotes the accuracy when using the baseline pre-trained model as the initial weights, and the dotted line represents the accuracy when using the fine-tuned baseline pre-trained model as the initial weights.

b The barplot present the accuracies of an ensemble of fine-tuned models, derived from the hybrid pre-trained models, for each of the datasets. The range of accuracy values indicated by the horizontal axis in each barplot is consistent with that in **a**.

a**b**

537

538 **Fig. 4: Grad-CAM images highlighting regions contributing to the disease classification**
 539 **by DL models for several test skin lesion images from the NSDD-labeled dataset.**

540 **a** Original skin lesion images of vitiligo and sebaceous conditions, paired with Grad-CAM images
541 using both the baseline pre-trained model and the self-supervised pre-trained model with BYOL.
542 Both disease labels were correctly classified. BYOL was selected for this analysis as it achieved the
543 highest accuracy among the three self-supervised learning methods.

544 **b** Original skin lesion images of lipoma and squamous cell carcinoma, paired with Grad-CAM images
545 using the same models as in **a**. Unlike **a**, the disease labels were misclassified in both cases.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)