Big Self-Supervised Models Advance Medical Image Classification

Supplementary Material

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan*, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, Mohammad Norouzi

Google Research and Health†

A. Datasets

A.1. Dermatology

Dermatology dataset details. As in actual clinical settings, the distribution of different skin conditions is heavily skewed in the Derm dataset, ranging from some skin conditions making up more than 10% of the training data like acne, eczema, and psoriasis, to those making up less than 1% like lentigo, melanoma, and stasis dermatitis [6]. To ensure that there was sufficient data to develop and evaluate the Dermatology skin condition classifier, we filtered the 419 conditions to the top 26 with the highest prevalence based on the training set. Specifically, this ensured that for each of these conditions, there were at least 100 cases in the training dataset. The remaining conditions were aggregated into an "Other" category (which comprised 21% of the cases in test dataset). The 26 target skin conditions are as follow: Acne, Actinic keratosis, Allergic contact dermatitis, Alopecia areata, Androgenetic alopecia, Basal cell carcinoma, Cyst, Eczema, Folliculitis, Hidradenitis, Lentigo, Melanocytic nevus, Melanoma, Post inflammatory hyperpigmentation, Psoriasis, Squamous cell carcinoma/squamous cell carcinoma insitu (SCC/SCCIS), Seborrheic keratosis, Scar condition, Seborrheic dermatitis, Skin tag, Stasis dermatitis, Tinea, Tinea versicolor, Urticaria, Verruca vulgaris, Vitiligo.

Figure A.1 shows examples of images in the Derm dataset. Figure A.2 shows examples of images belonging to the same patient which are taken from different viewpoints and/or from different body-parts under different lighting conditions. In the Multi Instance Contrastive Learning (MICLe) method, when multiple images of a medical condition from a given patient are available, we use two randomly selected images from all of the images that belong to this patient to directly create a positive pair of examples for contrastive learning.

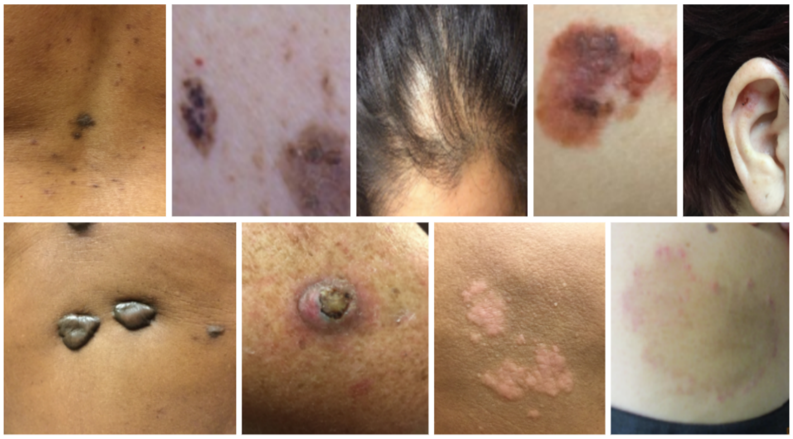

Figure A.1: Examples images from Derm dataset. Derm dataset includes 26 classes, ranging from skin conditions with greater than 10% prevalence like acne, eczema, and psoriasis, to those with sub-1% prevalence like lentigo, melanoma, and stasis dermatitis.

[*] Former intern at Google. Currently at Georgia Institute of Technology.

[†] {shekazizi, skornblith, iamtingchen, natviv, mnorouzi}@google.com

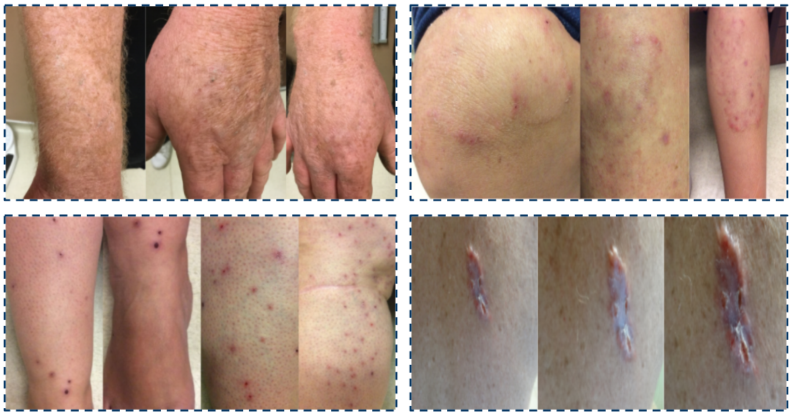

Figure A.2: Examples of images belong to the same patient which are taken from different viewpoints and/or from different body-parts under different lighting conditions. Each category, marked with a dashed line, belongs to a single patient and represents a single medical condition. In MICLe, when multiple images of a medical condition from the same patient are available, we use two randomly selected images from the patient to directly create a positive pair of examples and later adopt the augmentation. When a single image of a medical condition is available, we use standard data augmentation to generate two augmented views of the same image.

External dermatology dataset details. The dataset used for evaluating the out-of-distribution generalization performance of the model on the dermatology task was collected by a chain of skin cancer clinics in Australia and New Zealand. When compared to the in-distribution dermatology dataset, this dataset has a much higher prevalence of skin cancers such as Melanoma, Basal Cell Carcinoma, and Actinic Keratosis. It includes 8,563 de-identified multi-image cases which we use for the purpose of evaluating the generalization of the model under distribution shift.

A.2. CheXpert

Dataset split details. For CheXpert dataset [4] and the task of chest X-ray interpretation, we set up the learning task to diagnose five different thoracic pathologies: atelectasis, cardiomegaly, consolidation, edema and pleural effusion. The CheXpert dataset default split contains a training set of more than 200k images and a very small validation set that contains only 200 images. This extreme size difference is mainly because the training set is constructed using an algorithmic labeler based on the free text radiology reports while the validation set is manually labeled by board-certified radiologists. Similar to Neyshabur *et al.* [7, 8] findings, we realized due to the small size of the validation set, and the discrepancy between the label collection of the training set and the validation set, the high variance in studies is plausible. This variance implies that high performance on subsets of the training set would not correlate well with performance on the validation set, and consequently, complicating model selection from the hyper-parameter sweep. Following Neyshabur *et al.* [7] suggestion, in order to facilitate a robust comparison of our method to standard approaches, we define a custom subset of the training data as the validation set where we randomly re-split the full training set into 67,429 training images, 22,240 validation and 33,745 test images, respectively. This means the performances of our models are not compatible to those reported in [4] and the corresponding competition leader-board[1] for this specific dataset; nonetheless, we believe the relative performance of models is representative, informative, and comparable with [7, 8]. Figure A.3 shows examples of images in the CheXpert dataset which includes both frontal and lateral radiographs.

CheXpert data augmentation. Due to the less versatile nature of CheXpert dataset (see Fig. A.3), we used fairly strong data augmentation in order to prevent overfitting and improve final performance. At training time, the following preprocessing was applied: (1) random rotation by angle $\delta \sim U(-20, 20)$ degree, (2) random crop to 224×224 pixels, (3) random left-right flip with probability 50%, (4) linearly rescale value range from [0, 255] to [0, 1] followed by random additive brightness modulation and random multiplicative contrast modulation. Random additive brightness modulation adds a $\delta \sim U(-0.2, 0.2)$ to all channels. Random multiplicative contrast modulation multiplies per-channel standard deviation by a factor $s \sim U(-0.2, 0.2)$. After these steps we re-clip values to the range of [0, 1].

[1] https://stanfordmlgroup.github.io/competitions/chexpert/

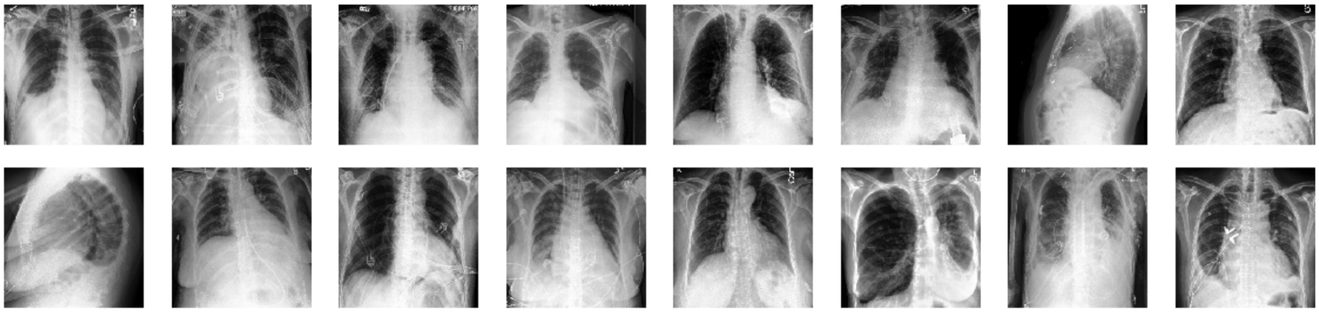

Figure A.3: Examples images from CheXpert dataset. The chest x-rays images are less diverse in comparison to the ImageNet and Derm dataset examples. The CheXpert task is to predict the probability of different observations from multi-view chest radiographs where we are looking for small local variations in examples using frontal and lateral radiographs.

B. Additional Results and Experiments

B.1. Dermatology Classification

B.1.1 Evaluation Details and Statistical Significance Testing

To evaluate the dermatology condition classification model performance, we compared its predicted differential diagnosis with the majority voted reference standard differential diagnosis (ground-truth label) using the top-k accuracy and the average top-k sensitivity. The top-k accuracy measures how frequently the top k predictions match any of the primary diagnoses in the ground truth. The top-k sensitivity measures this for each of the 26 conditions separately, whereas the final average top-k sensitivity is the average across the 26 conditions. Averaging across the 26 conditions avoids biasing towards more common conditions. We use both the top-1 and top-3 metrics in this paper.

In addition to our previous result comparing MICLe and SimCLR models against the supervised baselines, the non-parametric bootstrap is used to estimate the variability around model performance and investigating any significant improvement in the results using self-supervised pretrained models. Unlike the previous studies which uses confidence intervals obtained by multiple separate runs, for statistical significance testing, we select the best fine-tuned models for each of the architectures and compute the difference in top-1 and top-3 accuracies on bootstrap replicas of the test set. Given predictions of two models, we generate 1,000 bootstrap replicates of the test set and computing the difference in the target performance metric (top-k accuracy and AUCs) for both models after performing this randomization. This produces a distribution for each model and we use the 95% bootstrap percentile intervals to assess significance at the $p = 0.05$ level.

Table B.1 shows the comparison of the best self-supervised models *v.s.* supervised pretraining on dermatology classification. Our results suggest that, MICLe models can significantly ($p < 0.05$) outperform SimCLR counterpart and BiT [5] supervised model with ResNet-101 (3×) architecture over top-1 and top-3 accuracies. BiT model contains additional architectural tweaks included to boost transfer performance, and was trained on a significantly larger dataset of 14M images labelled with one or more of 21k classes which provides us with a strong supervised baseline *v.s.* the 1M images in ImageNet.

Table B.1: Comparison of the best self-supervised models *v.s.* supervised pretraining on dermatology classification. For the significance testing, we use bootstrapping to generate the confidence intervals. Our results show that the best MICLe model can significantly outperform BiT [5] which is a very strong supervised pretraining baseline trained on ImageNet-21k.

Architecture	Method	Top-1 Accuracy	Top-3 Accuracy
ResNet-152 (2×)	MICLe ImageNet→Derm (ours)	0.7037±0.0233	0.9273±0.0133
	SimCLR ImageNet→Derm [2]	0.6970±0.0243	0.9266±0.0135
ResNet-50 (4×)	MICLe ImageNet→Derm (ours)	0.7019±0.0224	0.9247±0.0135
	SimCLR ImageNet→Derm [2]	0.6975±0.0240	0.9271±0.0125
ResNet-101 (3×)	BiT Supervised [5]	0.6845±0.0228	0.9143±0.0142

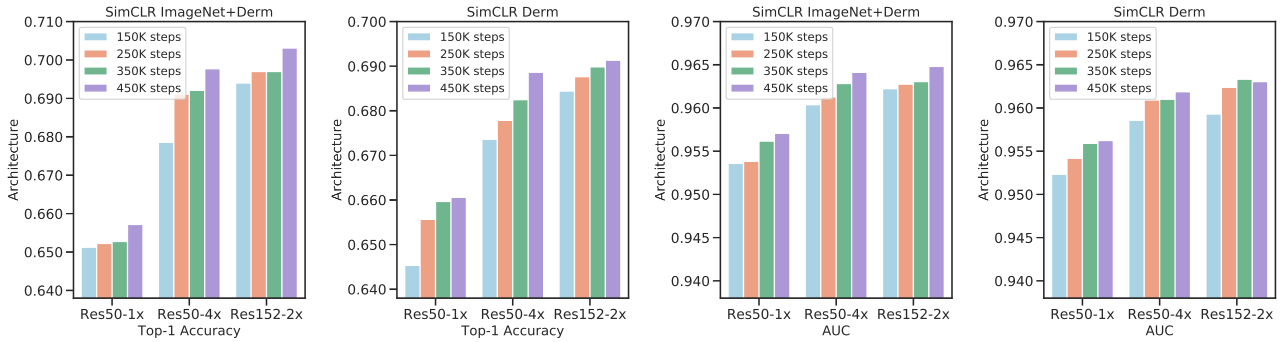

Figure B.4: Performance of dermatology condition classification models measured by the top-1 accuracy across different architecture and pretrained for 150,000 steps to 450,000 steps with a fixed batch size of 1024. Training longer provides more negative examples, improving the performance. Also, the results suggest that ImageNet initialization facilitating convergence, however, the performance gap between ImageNet initialized models and medical image only models are getting narrower.

B.1.2 Augmentation Selection for Multi-Instance Contrastive (MICLe) Method

To systematically study the impact of data augmentation in our multi-instance contrastive learning framework performance, we consider two augmentation scenarios: (1) performing standard simCLR augmentation which includes random color augmentation, crops with resize, Gaussian blur, and random flips, (2) performing a partial and lightweight augmentation based on random cropping and relying only on pair selections steps to create positive pairs. To understand the importance of augmentation composition in MICLe, we pretrain models under different augmentation and investigate the performance of fine-tuned models for the dermatology classification task. As the results in Table B.2 suggest, MICLe under partial augmentation often outperform the full augmentation, however, the difference is not significant. We leave comprehensive investigation of the optimal augmentations to future work.

Table B.2: Comparison of dermatology classification performance fine-tuned on representation learned using different unlabeled dataset with MICLe along with standard augmentation and partial augmentation. Our results suggest that MICLe under partial augmentation often outperform the full augmentation.

Architecture	Method	Augmentation	Top-1 Accuracy	Top-1 Sensitivity	AUC
ResNet-152 (2×)	MICLe Derm	Full Augmentation	0.6697	0.5060	0.9562
		Partial Augmentation	**0.6761**	0.5106	0.9562
	MICLe ImageNet→Derm	Full Augmentation	0.6928	0.5136	0.9634
		Partial Augmentation	0.6889	0.5300	0.9620
ResNet-50 (4×)	MICLe Derm	Full Augmentation	0.6803	0.5032	0.9608
		Partial Augmentation	**0.6808**	0.5204	0.9601
	MICLe ImageNet→Derm	Full Augmentation	0.6916	0.5159	0.9618
		Partial Augmentation	**0.6938**	0.5087	0.9629

B.1.3 Benefits of Longer Training

Figure B.4 shows the impact of longer training when models are pretrained for different numbers of epochs/steps. As suggested by Chen *et al.* [1, 3] training longer also provides more negative examples, improving the results. In this study we use a fixed batch size of 1024 and we find that with more training epochs/steps, the gaps between the performance of ImageNet initialized models with medical image only models are getting narrow, suggesting ImageNet initialization facilitating convergence where by taking fewer steps we can reach a given accuracy faster.

Furthermore, Fig. B.5 shows how the performance varies using the different available label fractions for dermatology task for the models pretrained for 150K steps and 450,000 steps using SimCLR ImageNet→Derm dataset. These results suggest that longer training yields proportionally larger gain for different label fractions. Also, this performance gain is more pronounced in ResNet-152 (2×). In fact, for ResNet-152 (2×) longer self supervised pretraining enable the model to match baseline using less than 20% of the training data *v.s.* 30% of the training data for 150,000 steps of pretraining.

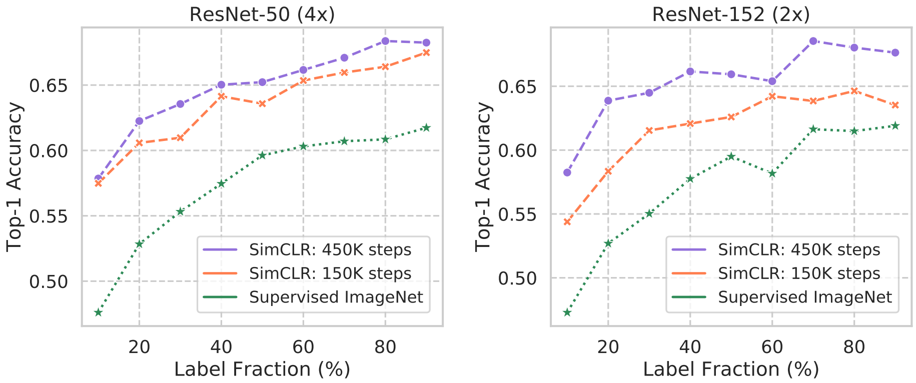

Figure B.5: Label efficiency progress over longer training for dermatology condition classification. The models are trained using ImageNet→Derm SimCLR for 150K steps and 450K steps and fine-tuned with varied sizes of label fractions. The Supervised ImageNet used as the baseline.

B.1.4 Detailed Performance Results

Table B.3 shows additional results for the performance of dermatology condition classification model measured by top-1 and top-3 accuracy, and area under the curve (AUC) across different architectures. Each model is fine-tuned using transfer learning from pretrained model on ImageNet, only unlabeled medical data, or pretrained using medical data initialized from ImageNet pretrained model. Again, we observe that bigger models yield better performance across accuracy, sensitivity and AUC for this task.

As shown in Table B.3, we once again observe that self-supervised pretraining with both ImageNet and in-domain Derm data is beneficial, outperforming self-supervised pretraining on ImageNet or Derm data alone. Moreover, comparing the performance of self-supervised models with Random and Supervised pretraining baseline, we observe self-supervised models significantly outperforms baselines ($p < 0.05$), even using smaller models such as ResNet-50 (1×).

Table B.4 shows additional dermatology condition classification performance for models fine-tuned on representations learned using different unlabeled datasets, and with and without multi instance contrastive learning (MICLe). Our results suggest that MICLe constantly improves the performance of skin condition classification over SimCLR [1, 3]. Using statistical significance test, we observe significant improvement for top-1 accuracy using MICLe for each dataset setting ($p < 0.05$).

Table B.3: Performance of dermatology condition classification models measured by top-1 and top-3 accuracy, and area under the curve (AUC) across different architectures. Models are pretrained for 150K steps and each model is fine-tuned using transfer learning from pretrained model on ImageNet, only unlabeled medical data, or pretrained using medical data initialized from ImageNet pretrained model. We observe that bigger models yield better performance.

Architecture	Method	Top-1 Accuracy	Top-3 Accuracy	Top-1 Sensitivity	Top-3 Sensitivity	AUC
ResNet-50 (1×)	SimCLR ImageNet	0.6258±0.0080	0.8943±0.0041	0.4524±0.0142	0.7388±0.0095	0.9480±0.0014
	SimCLR Derm	0.6249±0.0050	0.8967±0.0031	0.4402±0.0093	0.7370±0.0078	0.9485±0.0011
	SimCLR ImageNet→Derm	0.6344±0.0124	0.8996±0.0080	0.4554±0.0229	0.7349±0.0234	0.9511±0.0035
	Supervised ImageNet	0.5991±0.0174	0.8743±0.0094	0.4215±0.0267	0.7008±0.0225	0.9403±0.0044
	Random Initialization	0.5170±0.0062	0.8136±0.0108	0.3155±0.0152	0.5783±0.0031	0.9147±0.0019
ResNet-50 (4×)	SimCLR ImageNet	0.6462±0.0062	0.9082±0.0018	0.4738±0.0055	0.7614±0.0093	0.9545±0.0006
	SimCLR Derm	0.6693±0.0079	0.9173±0.0039	0.4954±0.0054	0.7822±0.0012	0.9576±0.0013
	SimCLR ImageNet→Derm	0.6761±0.0025	0.9176±0.0015	0.5028±0.0091	0.7828±0.0075	0.9593±0.0003
	Supervised ImageNet	0.6236±0.0032	0.8886±0.0024	0.4364±0.0096	0.7216±0.0070	0.9464±0.0005
	Random Initialization	0.5210±0.0177	0.8279±0.0172	0.3330±0.0203	0.6228±0.0314	0.9186±0.0060
ResNet-152 (2×)	SimCLR ImageNet	0.6638±0.0002	0.9109±0.0023	0.4993±0.0107	0.7716±0.0039	0.9573±0.0016
	SimCLR Derm	0.6643±0.0051	0.9126±0.0008	0.5035±0.0094	0.7808±0.0011	0.9558±0.0006
	SimCLR ImageNet→Derm	0.6830±0.0018	0.9196±0.0023	0.5156±0.0061	0.7891±0.0058	0.9620±0.0006
	Supervised ImageNet	0.6336±0.0012	0.8994±0.0022	0.4584±0.0162	0.7462±0.0076	0.9506±0.0015
	Random Initialization	0.5248±0.0121	0.8304±0.0127	0.3400±0.0303	0.6310±0.0366	0.9202±0.0055

Table B.4: Dermatology condition classification performance measured by top-1 accuracy, top-3 accuracy, and AUC. Models are fine-tuned on representations learned using different unlabeled datasets, and with and without multi instance contrastive learning (MICLe). Our results suggest that MICLe constantly improves the accuracy of skin condition classification over SimCLR.

Architecture	Method	Top-1 Accuracy	Top-3 Accuracy	Top-1 Sensitivity	Top-3 Sensitivity	AUC
ResNet-152 (2×)	MICLe Derm	0.6716±0.0031	0.9132±0.0022	0.5140±0.0093	0.7825±0.0027	0.9577±0.0009
	SimCLR Derm	0.6643±0.0051	0.9126±0.0008	0.5035±0.0094	0.7808±0.0011	0.9558±0.0006
	MICLe ImageNet→Derm	0.6843±0.0029	0.9246±0.0020	0.5199±0.0108	0.7933±0.0042	0.9629±0.0007
	SimCLR ImageNet→Derm	0.6830±0.0018	0.9196±0.0023	0.5156±0.0061	0.7891±0.0058	0.9620±0.0006
ResNet-50 (4×)	MICLe Derm	0.6755±0.0047	0.9152±0.0014	0.4900±0.0159	0.7603±0.0092	0.9583±0.0011
	SimCLR Derm	0.6693±0.0079	0.9173±0.0039	0.4954±0.0054	0.7822±0.0012	0.9576±0.0013
	MICLe ImageNet→Derm	0.6881±0.0036	0.9247±0.0011	0.5106±0.0076	0.7889±0.0091	0.9623±0.0005
	SimCLR ImageNet→Derm	0.6761±0.0025	0.9176±0.0015	0.5028±0.0091	0.7828±0.0075	0.9593±0.0003

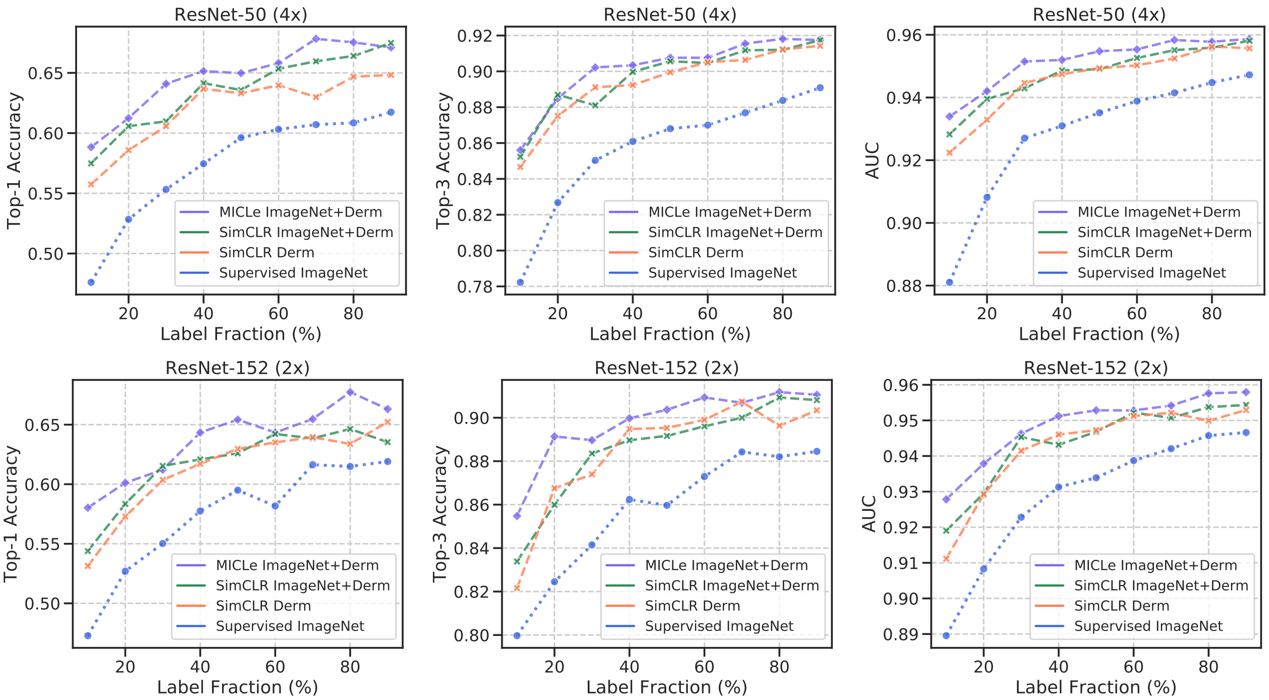

Figure B.6: The top-1 accuracy, top-3 accuracy, and AUC for dermatology condition classification for MICLe, SimCLR, and supervised models under different unlabeled pretraining dataset and varied sizes of label fractions. (top) ResNet-50 (4×), (bottom) ResNet-152 (2×).

B.1.5 Detailed Label Efficiency Results

Figure B.6 and Table B.5 provide additional performance results to investigate label-efficiency of the selected self-supervised models in the dermatology task. These results, back-up our finding that the pretraining using self-supervised models can significantly help with label efficiency for medical image classification, and in all of the fractions, self-supervised models outperform the supervised baseline. Also, we observe that MICLe yields proportionally larger gains when fine-tuning with fewer labeled examples and this is consistent across top-1 and top-3 accuracy and sensitivity, and AUCs for the dermatology classification task.

B.1.6 Subgroup Analysis

In another experiment, we also investigated whether the performance gains when using pretrained representations from self-supervised learning are evenly distributed across different subgroups of interest for the dermatology task; it is important for deployment in clinical settings that model performance is similar across such subgroups. We specifically explore top-1 and top-3 accuracy across different skin types of white, beige, brown, and dark brown. Figure B.7 shows the distribution

Table B.5: Classification accuracy and sensitivity for dermatology condition classification task, obtained by fine-tuning the SimCLR and MICLe on 10%, 50%, and 90% of the labeled data. As a reference, ResNet-50 (4×) fine-tuned the supervised ImageNet model and using 100% labels achieves 62.36% top-1 and 88.86% top-3 accuracy.

Performance Metric		Top-1 Accuracy			Top-3 Accuracy			Top-1 Sensitivity			Top-3 Sensitivity		
Architecture	Method	10%	50%	90%	10%	50%	90%	10%	50%	90%	10%	50%	90%
ResNet-152 (2×)	MICLe ImageNet→Derm	0.5802	0.6542	0.6631	0.8548	0.9037	0.9105	0.3839	0.4795	0.4947	0.6496	0.7567	0.7720
	SimCLR ImageNet→Derm	0.5439	0.6260	0.6353	0.8339	0.8916	0.9081	0.3446	0.4491	0.4786	0.6243	0.7269	0.7792
	SimCLR Derm	0.5313	0.6296	0.6522	0.8216	0.8953	0.9034	0.3201	0.4710	0.4906	0.6036	0.7373	0.7557
	Supervised ImageNet	0.4728	0.5950	0.6191	0.7997	0.8597	0.8845	0.2495	0.4303	0.4677	0.5452	0.7015	0.7326
ResNet-50 (4×)	MICLe ImageNet→Derm	0.5884	0.6498	0.6712	0.8560	0.9076	0.9174	0.3841	0.4878	0.5120	0.6555	0.7554	0.7771
	SimCLR ImageNet→Derm	0.5748	0.6358	0.6749	0.8523	0.9056	0.9174	0.3983	0.4889	0.5285	0.6585	0.7691	0.7902
	SimCLR Derm	0.5574	0.6331	0.6483	0.8466	0.8995	0.9142	0.3307	0.4387	0.4675	0.6233	0.7412	0.7728
	Supervised ImageNet	0.4760	0.5962	0.6174	0.7823	0.8680	0.8909	0.2529	0.4247	0.4677	0.5272	0.6925	0.7379

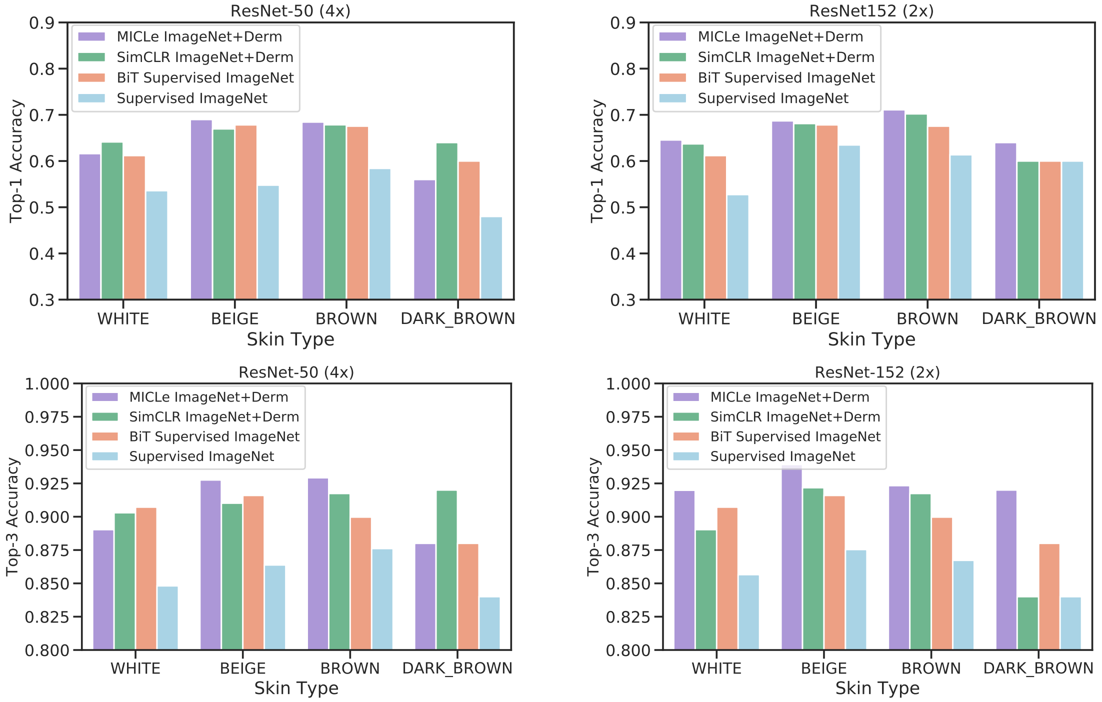

Figure B.7: Performance of the different models across different skin type subgroups for the dermatology classification task. Models pretrained using self-supervised learning perform much better on the rare skin type subgroups.

of performance across these subgroups. We observe that while the baseline supervised pretrained model performance drops on the rarer skin types, using self-supervised pretraining, the model performance is more even across the different skin types. This exploratory experiment suggests that the learnt representations are likely general and not picking up any spurious correlations during pretraining.

B.2. Chest X-ray Classification

B.2.1 Detailed Performance Results

For the task of X-ray interpretation on the CheXpert dataset, we set up the learning task to detect 5 different pathologies: atelectasis, cardiomegaly, consolidation, edema and pleural effusion. Table B.6 shows the AUC performance on the different pathologies on the CheXpert dataset. We once again observe that self-supervised pretraining with both ImageNet and in-domain medical data is beneficial, outperforming self-supervised pretraining on ImageNet or CheXpert alone. Also, the distribution of AUC performance across different pathologies suggests transfer learning, using both self-supervised and supervised models, provides mixed performance gains on this specific dataset. These observations are aligned with the findings of [8]. Although less pronounced, once again we observe that bigger models yield better performance.

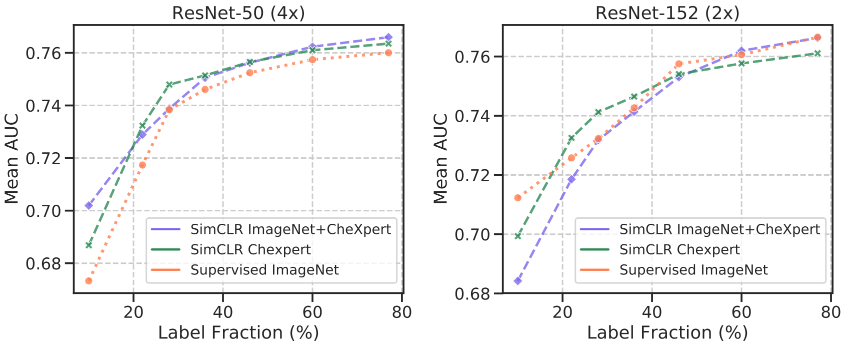

Figure B.8: Mean AUC for chest X-ray classification using self-supervised, and supervised pretrained models over varied sizes of label fractions for ResNet-50 (4×) and ResNet-152 (2×) architecture.

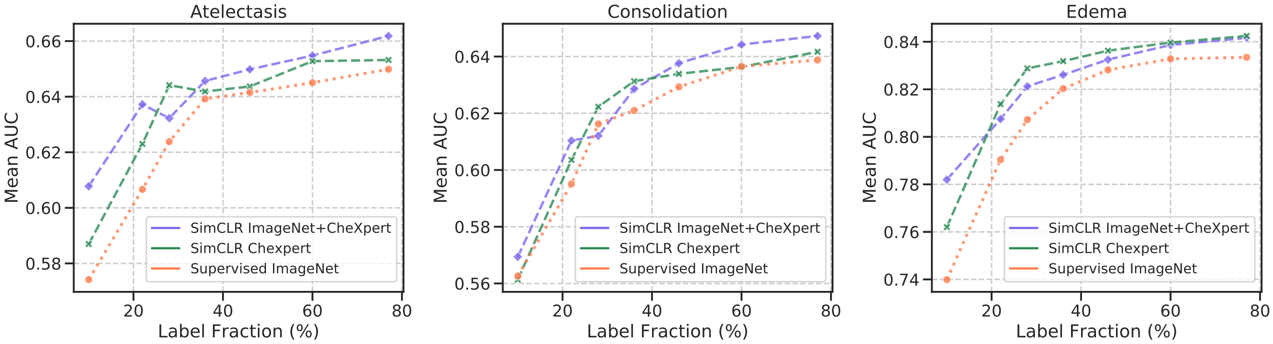

Figure B.9: Performances of diagnosing different pathologies on the CheXpert dataset measured with AUC over varied sizes of label fractions for ResNet-50 (4×).

Table B.6: Performances of diagnosing different pathologies on the CheXpert dataset measured with AUC. The distribution of AUC performance across different pathologies suggests transfer learning, using both self-supervised and supervised models, provides mixed performance gains on this specific dataset.

Architecture	Method	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
ResNet-50 (1×)	SimCLR ImageNet→CheXpert	0.6561±0.0052	0.8237±0.0024	0.6516±0.0051	0.8462±0.0008	0.8614±0.0016
	SimCLR CheXpert	0.6546±0.0030	0.8206±0.0025	0.6521±0.0027	0.8443±0.0012	0.8620±0.0005
	SimCLR ImageNet	0.6516±0.0046	0.8190±0.0015	0.6456±0.0036	0.8431±0.0012	0.8610±0.0010
	Supervised ImageNet	0.6555±0.0027	0.8188±0.0023	0.6517±0.0043	0.8429±0.0011	0.8607±0.0011
ResNet-50 (4×)	SimCLR ImageNet→CheXpert	0.6679±0.0022	0.8262±0.0026	0.6576±0.0039	0.8444±0.0012	0.8599±0.0018
	SimCLR CheXpert	0.6620±0.0038	0.8244±0.0017	0.6491±0.0029	0.8438±0.0014	0.8592±0.0013
	SimCLR ImageNet	0.6633±0.0025	0.8228±0.0014	0.6525±0.0028	0.8439±0.0015	0.8641±0.0013
	Supervised ImageNet	0.6570±0.0051	0.8218±0.0017	0.6546±0.0040	0.8425±0.0008	0.8624±0.0013
ResNet-152 (2×)	SimCLR ImageNet→CheXpert	0.6666±0.0027	0.8290±0.0019	0.6516±0.0024	0.8461±0.0016	0.8584±0.0015
	SimCLR CheXpert	0.6675±0.0040	0.8278±0.0015	0.6521±0.0030	0.8444±0.0013	0.8602±0.0016
	SimCLR ImageNet	0.6621±0.0067	0.8239±0.0014	0.6495±0.0046	0.8439±0.0013	0.8637±0.0014
	Supervised ImageNet	0.6496±0.0030	0.8224±0.0022	0.6498±0.0040	0.8408±0.0014	0.8615±0.0010

B.2.2 Detailed Label-efficiency Results

Figure B.8 and Fig. B.9 show how the performance changes when using different label fractions for the chest X-ray classification task. For architecture ResNet-50 (4×) self supervised models consistently outperform the supervised baseline, however, this trend is less striking for ResNet-152 (2×) models. We also observe that performance improvement in label efficiency is less pronounced for chest X-ray classification task in comparison to dermatology classification. We believe that with additional in-domain unlabeled data (we only use the CheXpert dataset for pretraining), self-supervised pretraining for chest X-ray classification improves.

References

[1] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019. 4, 5

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 4, 5

[4] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. 2

[5] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6, 2019. 3

[6] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, pages 1–9, 2020. 1

[7] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in Neural Information Processing Systems*, 33, 2020. 2

[8] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019. 2, 7