Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37469

Evaluating Pretraining Strategies for CNN-Based Skin Lesion Classification under Limited Labeled Data

Jingxuan Qu

College of Liberal Arts, University of Minnesota Twin Cities, Minneapolis, USA
qu000080@umn.edu

Abstract. Convolutional Neural Networks (CNNs) are widely used for image classification, but their performance is severely degraded when training data is limited. This paper studies the strategies to improve the performance and generalization of convolutional neural network under the condition of limited labels. Using the Human Against Machine with 10000 training images (HAM10000) skin lesion dataset, four training approaches are evaluated: supervised learning from scratch, ImageNet-based transfer learning, self-supervised learning, and a hybrid strategy that combines ImageNet pretraining with self-supervised adaptation to the target domain. Experimental results show that ImageNet transfer learning provides the strongest overall performance, demonstrating the value of pretrained representations under limited-data conditions. Self-supervised learning can exploit unlabeled target-domain images but does not consistently outperform conventional transfer learning. The proposed hybrid approach achieves performance close to the strongest ImageNet-based model, suggesting that domain-specific self-supervised adaptation can preserve much of the transferability of supervised pretrained features. In all, the results demonstrate the effectiveness of pretraining for limited data medical image classification and offer insight into the combination of supervised and self-supervised representation learning.

Keywords: convolutional neural networks, limited labeled data, transfer learning, self-supervised learning, skin lesion classification

1. Introduction

Convolutional neural networks are now commonly applied to automated skin lesion classification due to their strong performance in medical image analysis. Previous work has shown that CNNs can learn clinically useful visual patterns from both clinical and dermoscopic images, which has contributed to the development of computer-assisted diagnostic systems [1]. However, many successful systems depend on large, carefully labeled image collections. In real medical settings, image collection is costly and time-consuming, and accurate annotation relies on professional medical expertise. Institutional constraints, privacy issues and class imbalance may reduce the amount of data available for model training further [2]. Under such conditions, a CNN might fit well the available samples, but it might not generalize reliably to unseen cases. This problem is

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37469

particularly relevant in multi-class skin lesion classification, where high overall accuracy can hide weaker performance for infrequent classes.

Several strategies have been investigated to improve CNN learning when labeled data are scarce. The most common directions are data augmentation, regularization, reduced model complexity, transfer learning and self-supervised representation learning. All of these strategies are promising, but prior results do not favor any one as universally better. Brigato and Iocchi found that model complexity is a critical factor in small-data image classification and reported that lower-complexity CNNs, standard data augmentation, and dropout can remain highly effective when only a small number of samples are available [3]. In medical imaging, transfer learning is commonly adopted because pretrained models can provide a useful initialization; nevertheless, Raghu et al. showed that its benefit may be smaller than expected for some medical tasks and that lightweight models trained from scratch can perform comparably to large ImageNet-pretrained architectures [4]. Self-supervised learning provides another alternative by extracting informative features from unlabeled images before supervised fine-tuning, making it particularly relevant in settings where manual annotations are scarce [5].

The current evidence therefore suggests that the success of each strategy varies with factors, such as the dataset, class distribution, model architecture, domain similarity, and evaluation procedure. Skin lesion research has shown the feasibility of deep CNNs using large curated image collections [1], and public resources such as HAM10000 have supported the development and comparison of classification models [2]. However, comparing results across different studies is not straightforward. Differences in training-set size, data partitioning, preprocessing, model architecture, and evaluation choices can influence the reported results. For this reason, an improvement associated with data augmentation in one experiment should not be directly compared with a gain from transfer learning reported under a different setup. In addition, many studies mainly report the best final accuracy, while aspects such as the training-validation gap, performance stability on unseen samples, and variation across classes receive less attention.

This raises a practical research question that has not been fully addressed by previous work: under the same controlled limited-data setting, which training strategy offers the most robust improvement in CNN performance and generalization for multi-class skin lesion classification?

Therefore, under the same limited-label setting, this study compares four learning strategies: target-domain self-supervised learning, ImageNet-based transfer learning, supervised learning from random initialization, and a hybrid strategy that combines ImageNet pretraining with target-domain self-supervised adaptation. Every approach uses the same small amount of training data and follows the same validation and testing protocols. This enables a consistent comparison of the four strategies at various label scarcity levels.

This makes it easier to interpret the differences between the four strategies by evaluating them under the same limited-data setting. These findings can help in deciding which models and training strategies are better suited for skin lesion classification when labeled examples are scarce and might also be useful for other image classification tasks where annotation is costly or difficult to obtain. The comparison with standard ImageNet transfer learning also aims to assess whether target domain self-supervised adaptation is still useful even if the model is initialized with strong pre-trained features.

DOI: 10.54254/2755-2721/2026.37469

2. Related work

2.1. CNNs under limited-data conditions

The performance of CNNs in small data settings depends on the amount of training data and model complexity. Brigato and Iocchi experimented with different deep learning architectures on downsampled versions of the Canadian Institute for Advanced Research 10 (CIFAR-10), Fashion-MNIST and Street View House Numbers (SVHN) [3]. When only a few samples were available for each class, simpler CNNs often performed as well as, and in some cases better than, larger models. Standard data augmentation and dropout also remained useful under these conditions. Yao et al. reported a similar trend for skin lesion classification, where moderately complex models with regularization and modified augmentation strategy achieved competitive results on small and imbalanced dermoscopic datasets [6]. These results suggest that simply increasing the network depth or the number of parameters does not suffice to improve the performance with limited labeled data.

The results of these studies have to be understood in their own context as well. Brigato and Iocchi focused mainly on general-purpose benchmark datasets, so the same conclusions may not fully apply to fine-grained medical images [3]. Yao et al. examined small and imbalanced skin-lesion datasets, where limited data are closely connected with class imbalance and disease-specific visual characteristics [6]. Model complexity therefore cannot be considered independently of the dataset domain, class distribution, and the amount of labeled data available.

2.2. Data augmentation and regularization

In data-limited settings, augmentation and regularization are common means to improve model generalization. Augmentation creates greater variation among training examples through transformations, such as flipping, cropping, rotation and color jittering. Brigato and Iocchi demonstrated that even common augmentation can boost recognition performance in small-data scenarios significantly, and that dropout is still a valid regularization method [3]. Yao et al. applied the technique to skin lesion classification by combining Dropout, DropBlock, and Modified RandAugment with a moderately complex CNN [6]. Transfer learning-based methods have been successfully applied to classify dermoscopic skin lesions, providing additional evidence for the practical value of pretrained representations in this domain [7].

However, these approaches still rely mainly on better use of the labeled data at hand, rather than adding new semantic information. The effectiveness of such transformations can also depend on their type and intensity. This is especially relevant in medical imaging, as overly strong augmentation can alter patterns that are important for diagnosis, rather than adding harmless variation. Regularization is similar: too little and the model may memorize the training set, too much and it may not learn subtle class-specific features. Thus, augmentation and regularization are useful baselines but not a complete solution to the problem of learning robust representations from very limited labeled data.

2.3. Transfer learning

Transfer learning addresses the issue of limited labeled data by initializing a model with knowledge learned from a larger source dataset, most often ImageNet. This approach has become popular in medical imaging, as training a deep CNN from random initialization can be difficult when only a small number of annotations are available. However, Raghu et al. showed that ImageNet pretraining

DOI: 10.54254/2755-2721/2026.37469

does not always provide the expected benefits [4]. Across two medical imaging tasks, pretrained models sometimes produced only limited improvements, while relatively lightweight models could perform comparably to standard ImageNet architectures. Their analysis indicates that transfer learning should be evaluated empirically rather than assumed to be universally superior.

A major limitation of conventional transfer learning is the domain gap between natural and medical images. Features learned from ImageNet may not always capture the fine-grained textures and lesion-specific structures required for medical classification. In related work, the gap between ImageNet features and the properties of medical images has also been recognized, and transfer strategies have been explored using large collections of unlabeled medical images before fine-tuning on small labeled data [8]. This creates an important tension in the literature: external supervised pretraining is a convenient source of prior knowledge, but its usefulness may decrease as the source and target domains become more different.

2.4. Self-supervised learning

Self-supervised learning offers another approach to limited-label problems by first extracting representations from unlabeled images and then fine-tuning the model on a smaller labeled subset. Chen et al. introduced A Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [9], which builds a contrastive learning task from multiple augmented versions of the same image. The objective encourages features produced from a common source image to move closer in the representation space, while features originating from different images are separated. The study also emphasized the importance of augmentation design and showed that self-supervised representations can support effective downstream classification. In medical imaging, Zhou et al. reported that pretraining with unlabeled medical images can produce transferable features and improve downstream results compared with training from random initialization [10]. Similar self-supervised auxiliary strategies have also been explored for lightweight dermoscopic skin lesion classification, further indicating the usefulness of unlabeled representation learning in this area [11].

Self-supervised learning also has several practical limitations. To be successful, it needs a useful collection of unlabeled images, and the extra pretraining stage adds computational cost and implementation complexity. Contrastive methods such as SimCLR are very sensitive to the choice of augmentations, since the invariances learned by the models are determined by the transformations used to form positive pairs. In medical imaging, inappropriate transformations can eliminate or alter diagnostically important features. Thus, self-supervised learning can be beneficial when the amount of labeled data is limited, but its performance is dependent on the quality of unlabeled data, the design of the pretraining stage, and the relevance of the learned features to the downstream classification task.

2.5. Synthesis, research gap, and transition to methodology

A recurring tension appears in the literature on CNNs with limited data. Large-scale studies such as "Esteva et al. show that CNNs can reach good performance in skin-lesion classification when large amounts of labeled data, while real medical datasets such as HAM10000 are much smaller and expensive to annotate" [1]. This problem is addressed from different angles by existing solutions: augmentation and regularization make better use of available labels, transfer learning draws on knowledge from an external labeled source, and self-supervised learning learns representations from unlabeled data. However, no single approach is consistently superior across datasets and domains. Conventional augmentation may be insufficient when the labeled set is very

DOI: 10.54254/2755-2721/2026.37469

small; transfer learning may be affected by source-target domain mismatch; and self-supervised learning can reduce dependence on labels but requires additional unlabeled data and careful pretraining design.

These limitations motivate a controlled comparison of representation-learning strategies under the same limited-label conditions, including supervised learning from random initialization, ImageNet-based transfer learning, target-domain self-supervised learning, and a hybrid strategy combining ImageNet pretraining with target-domain self-supervised adaptation. Another question is whether target-domain self-supervised learning remains useful when it is applied not as an alternative to ImageNet pretraining, but as an adaptation stage after supervised pretraining. This motivates a direct comparison between conventional ImageNet transfer learning and the hybrid ImageNet-plus-target-domain self-supervised learning (SSL) strategy in this study.

Rather than assuming that one pretraining or representation-learning strategy will necessarily generalize better, the next section presents a methodology for evaluating how these strategies affect performance and generalization as labeled training data become scarce. The experimental design therefore uses consistent labeled-data budgets and evaluation procedures across all four strategies.

3. Methodology

This study investigates how different training strategies affect image classification performance under limited labeled-data regimes. Four methods are compared under the same data split, backbone architecture, and evaluation procedure: supervised learning from scratch, ImageNet-based transfer learning, target-domain self-supervised learning, and a hybrid approach combining ImageNet pretraining with target-domain SSL adaptation. The four methods differ primarily in how the feature representation is initialized and adapted before supervised classification.

3.1. Dataset and preprocessing

The experiments are carried out on the HAM10000 skin lesion dataset [2] which includes dermoscopic images of seven diagnostic categories. The dataset is suitable for the study as medical image classification is a practical problem where acquiring labeled data is costly and usually requires expert annotation.

All images are resized to 224 × 224 pixels before being fed to the models as input. The data are partitioned by lesion identity, not by individual images, to prevent images of the same lesion from appearing in both the training, validation, and test sets. The test set is held out during the development of the model and is only used for final performance evaluation.

Stratified sampling is used to split the lesion level training, validation and test sets, and to create the nested 10%, 25%, and 50% labeled subsets, to maintain the class distribution as close as possible across the splits.

3.2. Limited-label experimental setting

Only a portion of the training set's images are used as labeled examples in order to mimic situations where there are few labeled medical images. The experiments are performed with label budgets of 10%, 25% and 50% of the available training labels. The 10% setting represents the most extreme label-scarcity setting while the 25% and 50% settings are to examine the change of model performance as additional labeled data are available.

DOI: 10.54254/2755-2721/2026.37469

For each label budget, the same labeled subset is used in all four approaches. The remaining training images are not used with their labels in supervised training, but are available as unlabeled target domain data in Models C and D. This methodology allows the study to explore whether unlabeled photos can compensate for the lack of manual annotations.

3.3. Backbone architecture

The backbone architecture used in all experiments is ResNet50. Having the same backbone enables the comparison to be a matter of differences in representation initialization and pretraining strategies, not architectural differences. Depending on the training strategy used, the ResNet50 is used with either the random initialization or the ImageNet pretrained weights. In the supervised classification stage, the features extracted by the backbone are passed through the global average pooling and dropout before being fed into a seven-class softmax classifier for HAM10000.

3.4. Training strategies

The four training strategies are built on each other. Model A is the supervised baseline, Model B incorporates external ImageNet knowledge, Model C incorporates unlabeled target-domain information, and Model D incorporates both sources of prior knowledge.

3.4.1. Model A: CNN trained from scratch

Model A is the simple CNN baseline. The ResNet50 backbone is randomly initialized and trained directly with the only available labeled subset.

No external pretrained knowledge or unlabeled target-domain images are needed. Thus, the model must learn generic visual features along with class-specific representations from a limited number of labeled samples. This serves as a baseline to measure the benefits of transfer learning and self-supervised learning.

Supervised training is done with image augmentation to give the model more varied examples and less chance to overfit. The training images are randomly flipped horizontally and vertically, rotated and zoomed.

3.4.2. Model B: ImageNet pretraining with limited-label fine-tuning

Model B represents standard conventional transfer learning. The ResNet50 backbone is initialized using weights pretrained on ImageNet rather than random initialization.

The original ImageNet output layer is replaced by a new classifier configured for the seven HAM10000 categories. The model is then fine-tuned using the same limited labeled subset used by Model A.

This approach investigates whether ImageNet-derived visual knowledge can improve skin lesion classification under limited labeled-data conditions. However, because ImageNet and dermoscopic images have different visual distributions, the pretrained representations may not be fully adapted to lesion-specific patterns.

3.4.3. Model C: target-domain self-supervised pretraining

Model C evaluates whether useful representations can be learned directly from unlabeled target-domain images without relying on ImageNet pretraining.

DOI: 10.54254/2755-2721/2026.37469

The ResNet50 backbone is randomly initialized and first trained using self-supervised contrastive learning on unlabeled HAM10000 images from the training partition. A SimCLR-style strategy is adopted. For each image, two differently augmented views are generated to form a positive pair, while augmented views from different images are treated as negative pairs. The contrastive loss for a positive pair (i, j) is defined as:

$$\iota_{i,j} = -log\frac{\exp(\frac{sim(z_i,z_j)}{\tau})}{\sum_{k=1}^{2N} 1_{[k\neq i]}\exp(\frac{sim(z_i,z_j)}{\tau})} \tag{1}$$

In Equation (1), (z_i) and (z_j) denote the feature representations of two augmented views of the same image, (N) is the number of original images in a batch, and (τ) is the temperature parameter. The similarity function used in Equation (1) is defined using cosine similarity as follows:

$$sim(z_i, z_j) = \frac{z_i^T z_j}{\| z_i \| \|z_j\|} \tag{2}$$

In Equation (2), $z_i^T z_j$ denotes the inner product between the two feature vectors, while $\| z_i \|$ and $\|z_j\|$ denote their Euclidean norms.

By minimizing this objective, the model increases the similarity between representations of augmented views from the same image while distinguishing representations from different images, allowing useful features to be learned without diagnostic labels.

After self-supervised pretraining, the projection head is removed, and a skin lesion classifier is attached to the backbone. The backbone is then fine-tuned using the same limited labeled subset used by Models A and B.

The comparison between Models A and C therefore evaluates whether target-domain unlabeled images provide useful representations beyond supervised learning from scratch.

3.4.4. Model D: ImageNet pretraining with target-domain SSL adaptation

Model D combines transfer learning with self-supervised learning and represents the integrated strategy examined in this study.

Unlike Model C, the backbone does not begin with random weights. ResNet50 is first initialized with ImageNet pretrained weights and then further adapted to the HAM10000 image distribution through self-supervised contrastive learning using unlabeled training images. As in Model C, two augmented views of each image are generated and optimized using a SimCLR-style contrastive objective defined in Equation (1).

The projection head is eliminated and a classification head is added following target-domain SSL adaptation. Then, the model is fine-tuned on the same small labeled subset used by the other three methods.

The main difference between Models B and D is that the latter includes an extra target-domain adaptation step. Model B directly leverages ImageNet representations for the task of classification whereas Model D adapts these representations for the task first using unlabeled dermoscopic images. Therefore, Model D combines general visual knowledge from ImageNet and domain-specific knowledge from HAM10000. This design allows us to study if domain-specific self-

DOI: 10.54254/2755-2721/2026.37469

supervised adaptation provides any additional benefits over standard ImageNet transfer learning in the low-label regime.

3.5. Training procedure

For fair comparison, the same labeled training subsets, validation set and test set are used for all four methods. The same supervised data augmentation strategy is also applied consistently across models.

Model A is trained directly from random initialization by supervised learning. Models B, C and D were trained in two stages of supervised training. First, the encoder is frozen and only the classification head is trained. Then, the top 30 layers of the encoder are made trainable for fine-tuning, while batch normalization layers are frozen to improve training stability.

Models C and D use self-supervised learning only on photos from the training split. We omit validation and test pictures for the SSL pretraining to avoid information leaking.

The validation set is used for model selection and learning-rate adjustment. The test set is left untouched until the final evaluation . Early stopping, model checkpointing, and learning-rate reduction are applied based on validation loss.

3.6. Evaluation

The four methods are compared separately in terms of the test accuracy and the macro-averaged F1 score for each label budget. Accuracy summarizes overall correctness. Macro F1 averages class-wise F1 scores, giving equal weight to each class regardless of the number of examples. This makes macro F1 especially useful for HAM10000 where the class distribution is very imbalanced.

The analysis is not only about which method gives the best overall performance, but also about how the relative performance of the four strategies changes as the amount of labeled training data decreases. Specifically, the comparison of Models B and D is used to ascertain whether target-domain self-supervised adaptation offers advantages over traditional ImageNet transfer learning.

4. Experiments

4.1. Experimental setup

Experiments were conducted on the HAM10000 skin lesion dataset [2], which contains 10,015 dermoscopic images from seven diagnostic categories: actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc).

The dataset was split by lesion identity, to avoid leakage between related images. This ensured that all images associated to a given lesion were assigned to the same split. The resulting partition contained 8,017 training images, 1,029 validation images, and 969 test images. The validation and test partitions remained unchanged across all experiments.

To simulate different label-scarcity levels, three nested labeled subsets were built from the training set, containing approximately 10%, 25%, and 50% of the available labeled training images. These subsets contained approximately 802, 2,004, and 4,008 images respectively. For the sake of fair comparison, all models used the same data subsets.

ResNet50 is adopted as the backbone architecture with the input resolution of 224 × 224 for all the experiments. All models were trained with the same data augmentation strategy for supervised

DOI: 10.54254/2755-2721/2026.37469

training which included random horizontal and vertical flipping, random rotation and random zoom. A global average pooling layer followed by a dropout layer (rate = 0.3) was appended prior to the seven-way softmax classification head.

Four training strategies were compared:

• Model A: ResNet50 initialized randomly and trained using supervised learning only.

• Model B: ResNet50 initialized with ImageNet-pretrained weights and subsequently fine-tuned on the labeled HAM10000 subsets.

• Model C: ResNet50 initialized randomly and first pretrained using self-supervised contrastive learning on all images from the training partition, followed by supervised fine-tuning.

• Model D: ResNet50 initialized with ImageNet-pretrained weights, further adapted to the HAM10000 domain using self-supervised contrastive learning, and finally fine-tuned using labeled data.

For Models C and D, self-supervised learning was performed only on images from the training partition. Labels were not used during this stage, and no validation or test images were included. The same self-supervised encoder was reused for the 10%, 25%, and 50% labeled-data experiments.

For self-supervised learning, two augmented versions of each image were generated and optimized using a contrastive objective. The augmentations included random flipping, rotation, zoom, contrast adjustment, and brightness adjustment. Adam was used for optimization. Model C was pretrained for a maximum of 10 epochs with a learning rate of 1×10^{-4}, whereas Model D used a smaller learning rate of 1×10^{-5} during the 10-epoch self-supervised adaptation stage to reduce disruption to the pretrained ImageNet representations.

A batch size of 16 was maintained throughout all experiments, with no explicit weight decay. Model A used Adam with an initial learning rate of 1×10^{-4} for a maximum of 30 epochs. For Models B, C, and D, supervised training was conducted in two stages. First, the encoder remained fixed, and only the classification layers were optimized for up to 10 epochs using Adam with a learning rate of 1×10^{-3}. The upper 30 layers of the encoder were then made available for fine-tuning, while batch normalization layers remained frozen. Fine-tuning continued for up to 20 epochs at 1×10^{-5}. Validation loss was used to guide early stopping, checkpoint selection, and learning-rate reduction.

4.2. Evaluation metrics

Because HAM10000 is strongly imbalanced, overall classification accuracy alone may overemphasize performance on the majority class. Therefore, both accuracy and macro-averaged F1 score were used as the primary evaluation metrics.

Accuracy was calculated as the ratio of correct predictions to the total number of test samples:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{3}$$

In Equation (3), the numerator represents the number of correctly classified test images and the denominator represents the total number of test images.

Macro F1 was calculated by computing the F1 score independently for each of the seven classes and then averaging them with equal weight:

DOI: 10.54254/2755-2721/2026.37469

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C} F1_i \tag{4}$$

In Equation (4), C represents the total number of lesion categories, which is seven for HAM10000, and $F1_i$ denotes the F1 score calculated for the i -th class. Macro F1 is more robust to class imbalance and effectively reflects model generalization on both dominant and rare lesion subtypes.

4.3. Experimental results

Table 1 summarizes test accuracy and macro F1 across the 10%, 25%, and 50% labeled-data settings. Across the experiments, pretrained representations generally provided stronger performance than training from random initialization, especially under the smaller label budgets. The strongest overall performance was obtained using ImageNet-pretrained ResNet50, while the proposed hybrid strategy yielded stable and competitive results across all data sparsity levels.

Table 1. Performance comparison of the four models under different labeled-data ratios

Model	10% Accuracy	10% Macro F1	25% Accuracy	25% Macro F1	50% Accuracy	50% Macro F1
Model A: Supervised from Scratch	0.7018	0.3523	0.7079	0.3761	0.7214	0.4181
Model B: ImageNet Transfer Learning	0.7296	0.4019	0.7781	0.6176	0.7812	0.5654
Model C: Target-domain SSL	0.6852	0.3503	0.7069	0.3833	0.7255	0.4211
Model D: ImageNet + Target-domain SSL	0.7265	0.3964	0.7528	0.5191	0.7694	0.5463

The results in Table 1 confirm the importance of large scale pretraining in the limited data setting. With only 10% of labeled training data, the accuracy and macro F1 of ImageNet-based transfer learning reached 0.7296 and 0.4019 respectively, which was significantly higher than random initialization. This benefit is more pronounced in the 25% label settings with macro F1 score of 0.6176.

The self-supervised pretraining on the target domain from random initialization was more gradual with more labeled data. Accuracy went up from 0.6852 at 10% labels to 0.7255 at 50%. Macro F1 score went up from 0.3503 to 0.4211. These results indicate that self-supervised learning on HAM10000 learned useful domain-specific representations, but its representation ability is limited when compared with large-scale general domain pretraining. This distinction is particularly relevant to skin lesion classification, since some clinically meaningful lesion categories are far less prevalent than melanocytic nevi in HAM10000. Thus, the improvement of the overall accuracy does not mean that the recognition of all lesion types is equally well.

The proposed approach follows ImageNet initialization, followed by further self-supervised adaptation on unlabeled HAM10000 training images, before supervised fine-tuning. This strategy reached the accuracy of 0.7265, 0.7528 and 0.7694 in the label settings of 10%, 25% and 50% respectively, with the macro F1 scores of 0.3964, 0.5191 and 0.5463.

DOI: 10.54254/2755-2721/2026.37469

The proposed method did not beat the standard ImageNet transfer learning, but it performed consistently close to the strongest baseline. For example, for the label setting of 10%, the accuracy difference between the proposed method and ImageNet transfer learning was only 0.31 percentage points, and the macro F1 difference was 0.55 percentage points. The proposed method attained an accuracy of 0.7694 at 50% label level as compared to the ImageNet transfer learning of 0.7812.

This is a significant result as the additional self-supervised adaptation stage changes the pretrained representation significantly without any additional class labels. The fact that the performance was close to the ImageNet baseline indicates that the domain-adaptation procedure was able to preserve much of the useful general visual knowledge learned from ImageNet, while adapting the encoder to the visual characteristics of dermoscopic images.

However, the fact that no significant improvement over standard transfer learning is achieved also shows that domain-specific self-supervised adaptation does not necessarily improve downstream classification results. One possible explanation is that ImageNet pre-training already provides sufficiently robust and transferable generic features for HAM10000. Thus, any extra contrastive adaptation can only give limited new information, while possibly changing some pretrained features that were already useful for classification.

The different behavior of Models C and D further illustrates how the effect of target-domain self-supervised learning depends on the starting representation. Self-supervised pretraining provides useful structure for Model C that would not be present with random initialization, allowing the model to learn domain-specific visual patterns prior to supervised fine-tuning. Model D, in contrast, starts with strong ImageNet representations. In this case, the extra SSL stage has less new information to add, and can also partially change the pretrained features that are already beneficial for lesion classification. Thus, target domain SSL is relatively more beneficial when starting from random initialization, and less beneficial (or even slightly negative) when used after strong ImageNet pretraining.

4.4. Discussion

Experimental results demonstrate that large-scale supervised pretraining is still very effective for skin lesion classification when the labeled data is limited. Results with ImageNet initialization consistently generated strong representations that transferred well to HAM10000, providing evidence that general visual features learned from a large and diverse source dataset can remain useful even when dermoscopic medical images are very different from natural scene images.

The proposed hybrid strategy achieves a robust and balanced trade-off.. By starting from ImageNet-pretrained weights and then applying self-supervised adaptation to HAM10000, the model retained performance close to that of the strongest ImageNet-only baseline. This suggests that target-domain self-supervised adaptation can modify pretrained representations without substantially reducing their downstream usefulness. But the absence of a clear gain over standard transfer learning also suggests that further domain adaptation may have diminishing returns when the initial pretrained representation is already strong.

There could be several reasons why there isn't a noticeable improvement over conventional ImageNet transfer learning. First, ImageNet pretraining provides a strong and transferable visual representation for HAM10000,leaving relatively limited room for improvement. Further contrastive adaptation may adjust some of the already useful pretrained features to suit downstream classification. Second, despite the availability of thousands of unlabeled training images from HAM10000, the scale of HAM10000 is still small compared to the datasets used in successful self-supervised representation learning. Third, augmentations for contrastive learning might not be fully

DOI: 10.54254/2755-2721/2026.37469

consistent with dermoscopic diagnosis. Skin lesion classification can be performed based on subtle visual characteristics such as lesion shape, boundary irregularity, pigmentation patterns and local texture, some of which can be altered during augmentation. Finally, the adaptation learning rate and training schedule may not provide the best trade-off between preserving ImageNet knowledge and learning target-domain features. A more carefully tuned adaptation strategy may be needed, therefore, to obtain further improvements.

Class imbalance is another important factor. HAM10000 has a substantially larger number of images of melanocytic nevi compared to several minority lesion types. Consequently, accuracy can be relatively high, but the performance on the minority class is poor. The differences between accuracy and macro F1 observed illustrate the importance of evaluating both metrics. The Macro F1 is particularly useful in providing a more balanced evaluation across the lesion categories.

To mitigate the influence of stochastic training variation, all four models were evaluated over multiple runs with differing random seeds, and the reported results were averaged by mean performance across runs. There was some variability but the overall trend of the ranking and performance was similar. This means that the experimental results are rather robust. Additional runs would increase the statistical reliability of the comparison however.

Overall, the results show that ImageNet pretraining is still the strongest and most robust method in this low-label setting, and the proposed hybrid approach shows that large-scale pretrained representations can be further adapted to the target domain through self-supervised learning without significant loss of downstream performance. Thus, combining supervised pretraining and target-domain self-supervision is a promising direction, but further gains may require more carefully designed contrastive objectives, augmentation strategies, or larger amounts of unlabeled target-domain data.

5. Conclusion

This study examined CNN performance and generalization under restricted labeled-data conditions. Using the HAM10000 skin lesion dataset, four training strategies were evaluated, including supervised learning from scratch, ImageNet-based transfer learning, self-supervised learning, and a hybrid approach combining ImageNet pretraining with self-supervised target-domain adaptation.

The experimental results demonstrate that the choice of training and representation initialization strategy has a substantial influence on classification performance under limited-data conditions. Weaker performance was generally observed with training from scratch, as both generic visual features and task-specific patterns had to be learned from a relatively small labeled set. Conversely, transfer learning from ImageNet obtained the best overall performance, suggesting that features learned from large natural-image datasets may still be useful for dermoscopic classification.

Self-supervised learning was an approach to leverage unlabeled images from the target domain, but the downstream performance was not always better than conventional transfer learning. The proposed hybrid approach achieves performance close to the strongest ImageNet based model. The result is not a clear advantage over standard transfer learning, but does indicate that the self-supervised target-domain adaptation in the proposed method can change the pretrained representations while still maintaining much of their transferability in the original domain.

To summarize, the results demonstrate the practical value of pretrained representations for medical image classification in the setting of limited labeled data. Among the methods considered the classical ImageNet transfer learning is still the best and most useful approach. The hybrid method shows that large-scale supervised pretraining can be followed by domain-specific self-supervised adaptation without significant performance loss.

DOI: 10.54254/2755-2721/2026.37469

The current experiments can be extended in several ways in future work. Other medical image datasets and different proportions of labeled data could be examined to see whether the trends observed here extend beyond the HAM10000 dataset. Tougher label-scarcity scenarios could also illuminate situations where self-supervised learning is more beneficial.

Other self-supervised methods could be explored such as more advanced contrastive approaches and masked-image modeling approaches. In addition, the duration of adaptation, design of augmentation and configuration of fine-tuning might also impact the performance of the hybrid strategy.

Class imbalance is also worth further investigation. HAM10000 has a large class frequency imbalance across the diagnostic categories, which can negatively impact the performance on the minority classes. Therefore, methods such as class-balanced sampling, weighted loss functions, and focal loss could be explored to improve macro-level performance and to achieve a more consistent classification across lesion categories.

References

[1] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*, 115–118.

[2] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data, 5*, 180161.

[3] Brigato, L., & Iocchi, L. (2021). A close look at deep learning with small data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2490–2497. IEEE. https://doi.org/10.1109/ICPR48806.2021.9412492

[4] Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems, 32*, 3342–3352.

[5] Krammer, S., Li, Y., Jakob, N., Boehm, A. S., Wolff, H., Tang, P., Lasser, T., French, L. E., & Hartmann, D. (2022). Deep learning-based classification of dermatological lesions given a limited amount of labelled data. *Journal of the European Academy of Dermatology and Venereology, 36*(12), 2516–2524.

[6] Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., Shao, P., Kaffenberger, B., & Xu, R. X. (2022). Single model deep learning on imbalanced small datasets for skin lesion classification. *IEEE Transactions on Medical Imaging, 41*(5), 1242–1254. https://doi.org/10.1109/TMI.2021.3136682

[7] Fraiwan, M., & Faouri, E. (2022). On the automatic detection and classification of skin cancer using deep transfer learning. *Sensors, 22*(13), 4963.

[8] Alzubaidi, L., Al-Amidie, M., Al-Asadi, A., Humaidi, A. J., Al-Shamma, O., Fadhel, M. A., Zhang, J., Santamaría, J., & Duan, Y. (2021). Novel transfer learning approach for medical imaging with limited labeled data. *Cancers, 13*(7), 1590.

[9] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, 119*, 1597–1607.

[10] Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., & Liang, J. (2021). Models genesis. *Medical Image Analysis, 67*, 101840. https://doi.org/10.1016/j.media.2020.101840

[11] Wang, Y., Wang, Y., Cai, J., Lee, T. K., Miao, C., & Wang, Z. J. (2023). SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. *Medical Image Analysis, 84*, 102693.