Meta-DermDiagnosis: Few-Shot Skin Disease Identification using Meta-Learning

Kushagra Mahajan, Monika Sharma, Lovekesh Vig
TCS Research, New Delhi, India
Email:{kushagra.mahajan, monika.sharma1, lovekesh.vig}@tcs.com

Abstract

Annotated images for diagnosis of rare or novel diseases are likely to remain scarce due to small affected patient population and limited clinical expertise to annotate images. Deep networks employed for image based diagnosis need to be robust enough to quickly adapt to novel diseases with few annotated images. Further, in case of the frequently occurring long-tailed class distributions in skin lesion and other disease classification datasets, conventional training approaches lead to poor generalization on classes at the tail end of the distribution due to biased class priors. This paper focuses on the problems of disease identification and quick model adaptation in such data-scarce and long-tailed class distribution scenarios by exploiting recent advances in meta-learning. This involves training a neural network on few-shot image classification tasks based on an initial set of class labels / head classes of the distribution, prior to adapting the model for classification on a set of unseen / tail classes. We named the proposed method Meta-DermDiagnosis because it utilizes meta-learning based few-shot learning techniques such as the gradient based Reptile and distance metric based Prototypical networks for identification of diseases in skin lesion datasets. We evaluate the effectiveness of our approach on publicly available skin lesion datasets, namely the ISIC 2018, Derm7pt and SD-198 datasets and obtain significant performance improvement over pre-trained models with just a few annotated examples. Further, we incorporate Group Equivariant convolutions (G-convolutions) for the Meta-DermDiagnosis network to improve disease identification performance as these images generally do not have any prevailing global orientation / canonical structure and G-convolutions make the network equivariant to any discrete transformations like rotation, reflection and translation.

1. Introduction

Over the past decade, the availability of large quantities of labeled data has enabled deep learning methods to achieve impressive breakthroughs in learning tasks such as speech recognition, object recognition and machine translation. Additionally, deep learning has also proven its value in the automation of medical image analysis to potentially assist doctors in the effective diagnosis and treatment of diseases such as detection of breast cancer from mammograms [38, 29], tumors from CT scan images [16, 12, 33], and pathologies from chest X-rays [34, 7].

Another demanding medical specialization that stands to benefit from deep learning is Dermatology with cases of skin diseases outpacing hypertension, obesity and cancer summed together. Automated classification of skin lesions using images is a particularly challenging task owing to the long-tailed class distribution of skin datasets (shown in Figure 1), fine-grained variability in the appearance of skin lesions, and the lack of sufficient images available for the novel skin ailments being discovered. Annotations of these skin diseases is very time consuming, labour intensive, costly and error-prone even when it is performed by experienced doctors. This motivated researchers to apply deep models for automated diagnoses. However, these networks tend to fail when there is limited annotated data available since they over-fit and are less likely to generalize well. Moreover, these methods learn skewed class priors towards dominant classes of the distribution and do not generalize to tail classes in case of heavy-tailed class distributions. In contrast, humans can learn quickly from a few examples by leveraging prior knowledge. Such capacity in data efficiency and fast adaptation, if realized in machine learning, can greatly expand its utility.

To circumvent the issue of scarce data / heavy-tailed class distributions, methods for few shot classification such as transfer learning [6, 30, 11] were proposed. However, these methods are successful only when sufficient labeled data is available in the target domain and do not guarantee optimal network initialization parameters that can quickly adapt to new target domains. Hence, to facilitate learning from small amounts of annotated data, meta-learning [25] techniques have emerged. These techniques imbibe the system with the capability to rapidly adapt to new tasks and environments with very few training examples. The key un-

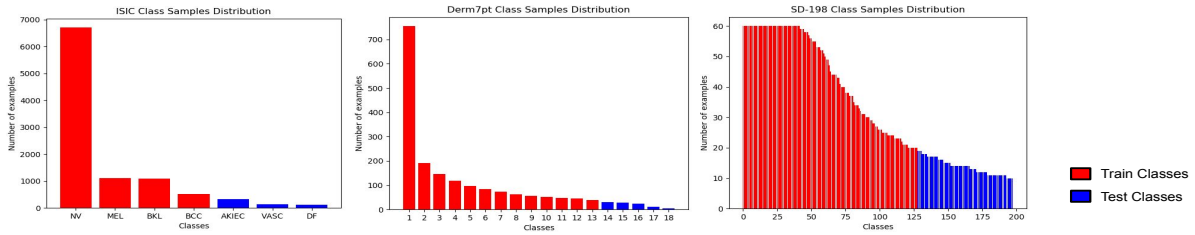


Figure 1. Figures showing class distribution in skin lesion datasets: ISIC 2018 [4], Derm7pt [13] and SD-198 [32]. The distribution is generally heavy-tailed with some classes having very few samples.The classes towards head of the class distribution (common-diseases), shown in **red**, are taken as train classes and classes at the tail of the distribution (new / rare disease), shown in **blue** color, are chosen as test classes.

derlying idea is to train the model's initial parameters such that the model has maximal performance on a new task after the parameters have been updated through one or more gradient steps computed with a small amount of data from the new task.

In this paper, we explore meta-learning based few-shot approaches like Reptile [25] (i.e., gradient-based method) and Prototypical networks [31] (i.e., distance metric based learner) to identify skin lesions from medical images in low-data / heavy-tailed data distribution regimes. We call the proposed network *Meta-DermDiagnosis* which utilizes meta-learning to facilitate quick adaptation of deep neural networks trained on data samples of common diseases for identification of rare diseases with much less annotated data. In essence, it consists of a meta-learner which involves training the neural network to solve a number of few-shot image classification tasks based on an initial set of class labels. The class labels are sampled from the head of the class distribution to find effective network initialization weights and subsequently, adapting the model to perform classification on a new set of unseen classes / tail classes with very few examples. Furthermore, we demonstrate that using Group Equivariant convolutions (G-convolutions) [5] in Meta-DermDiagnosis greatly improves the network's performance in case of skin lesion image classification as orientation is generally not an important feature in such images. We evaluate the performance of our proposed method using Reptile and Prototypical networks on three publicly available skin lesion classification datasets namely, the ISIC 2018 [4], Derm7pt [13] and SD-198 [32], and compare their performance against the pre-trained transfer learning baseline. Our results demonstrate that Reptile with G-convolutions performs better than other approaches in the low-data regime. We can also use the proposed approach for disease identification in other medical imaging datasets. As a summary, we make following contributions in the paper:

- We propose to use meta-learning for rare disease identification in skin lesion image datasets having long-tailed class distributions and few annotated data samples by formulating the problem as 1-shot, 3-shot and 5-shot classification problems. We named the proposed network *Meta-DermDiagnosis*.
- We explore the gradient based Reptile [25] and metric-learning based Prototypical networks [31] for identifying diseases from skin lesion images in low-data regimes and present the results in Section 5.3.
- Further, we demonstrate that G-convolutions [5] greatly improve the network's performance in case of skin lesion images as orientation is generally not an important feature in such skin lesion data.
- We evaluate Meta-DermDiagnosis network on publicly available skin lesion datasets such as ISIC 2018 [4] , Derm7pt [13] and SD-198 [32] and compare the classification performance with pre-training as a baseline, as described in Section 5.3. The results demonstrate that Reptile outperforms pre-training and prototypical networks in skin lesion classification in low-data scenarios.
- We also claim that the proposed meta-learning based disease identification system can also be applied on other medical imaging datasets in future work.

The remaining paper is organized as follows : In Section 2, we contrast our work to the existing literature. Section 3 defines the problem of few-shot learning for skin-disease identification and outlines the meta-learning techniques employed. Subsequently, we provide details of data sets used in Section 4 and present the results of the experiments conducted and their discussion in Section 5. Conclusions are presented in Section 6.

2. Related Work

Deep neural networks (DNNs) are state-of-the-art models across various domains ranging from image analysis [10, 22, 24] to natural language processing [18, 2]. Recently, deep learning has also shown great success in medical image classification and segmentation [9, 21, 1, 3, 39, 26, 8, 34, 7]. However, DNNs are most effective when large

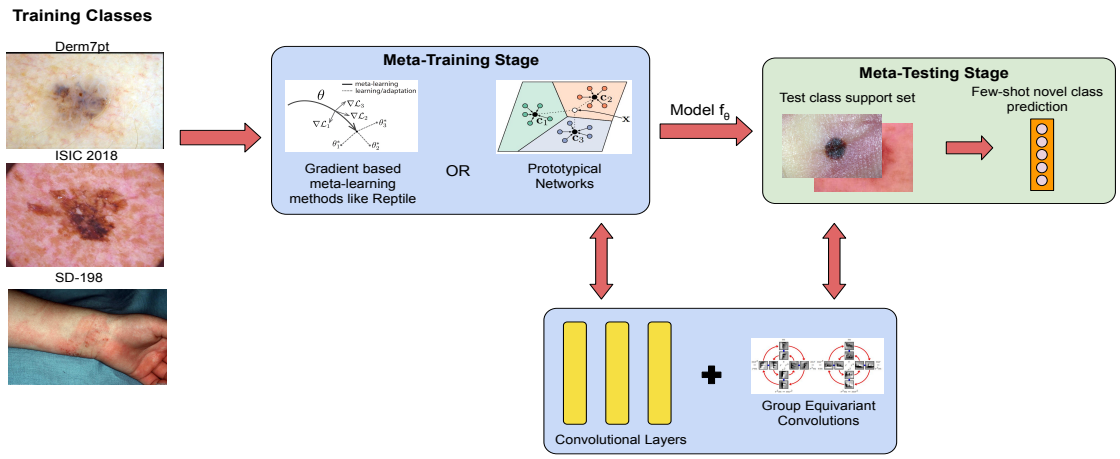


Figure 2. Figure showing an overview of the proposed approach *Meta-DermDiagnosis* for identification of diseases in skin lesion datasets based on meta-learning techniques Reptile and Prototypical networks.

volumes of annotated data is available for training which is generally not the case with medical data of rare diseases.

Several approaches to deal with scarce data have been proposed in the literature such as transfer learning [28, 11, 30, 6] and few-shot learning [15, 36] which requires the network to be pre-trained on a large amount of labelled data on a related domain and subsequently, fine-tuned on target domain data. While meta-learning techniques have been successfully applied for classifying real world image datasets, their application to medical images has been very limited [35, 23, 25, 31, 37]. Skin lesion classification is one such medical field where researchers have not sufficiently explored meta-learning techniques. However, we have demonstrated that meta-learning techniques prove to be beneficial in skin lesion classification datasets because of limited availability of samples for a large number of existing or new skin diseases. Prabhu *et al*. [27] proposed learning a mixture of prototypes for each disease class initialized via clustering and refined via an online update scheme. In another paper [20], authors have proposed a difficulty-aware meta-learning method that dynamically monitors the importance of learning tasks during the meta-optimization stage and evaluate their network's performance on ISIC 2018 skin lesion dataset. In our work, we utilized meta learning techniques such as Reptile and Prototypical networks, along with G-convolutions for skin disease classification and significantly outperformed DAML [20] on the ISIC 2018 [4] skin lesion dataset. In addition, we also performed experiments on other skin lesion datasets like Derm7pt [13] and SD-198 [32] and report encouraging results.

3. Meta-DermDiagnosis

The objective of the paper is to identify diseases from skin lesion images. The image datasets for skin lesions generally have skewed distribution among different lesion classes i.e., a long-tailed distribution as shown in Figure 1. This is due to the fact that new/rare diseases are being discovered everyday which are difficult to annotate because of limited expertise. This limits the number of annotated samples for new / rare diseases as compared to those of common diseases. If we train conventional deep networks to classify the skewed skin lesion datasets, they do not generalize well as they learn biased class priors towards the classes with larger number of samples and give poor performance on rare disease classes. Moreover, repeated training of deep networks is a time-consuming process, which is often not desirable in healthcare. This motivates the exploration of techniques which can quickly learn and adapt to new disease classes with very few annotated examples.

Therefore, we formulate the problem of disease identification from skin lesion images in low-data regimes as a few-shot learning problem by utilizing recent meta-learning techniques. We call the proposed solution *Meta-DermDiagnosis* that aims to facilitate quick adaptation of networks trained on common diseases to identification of new / rare disease classes with limited annotated data. An overview of the entire pipeline for Meta-DermDiagnosis is shown in Figure 2 which includes a meta-training stage and a meta-testing stage. The meta-training stage consists of a meta-learner for training the neural network to solve a large number of few-shot image classification tasks created from a set of training classes comprising of common diseases, with the classes being sampled from the head of the distribution, and finding effective network initialization parameters for the model. The meta-learning technique can either be a gradient based Reptile algorithm or a distance-metric based Prototypical network. Next, in the meta-testing stage, the model is adapted to perform classification on a new set of unseen / rare classes with very few examples. We also uti-

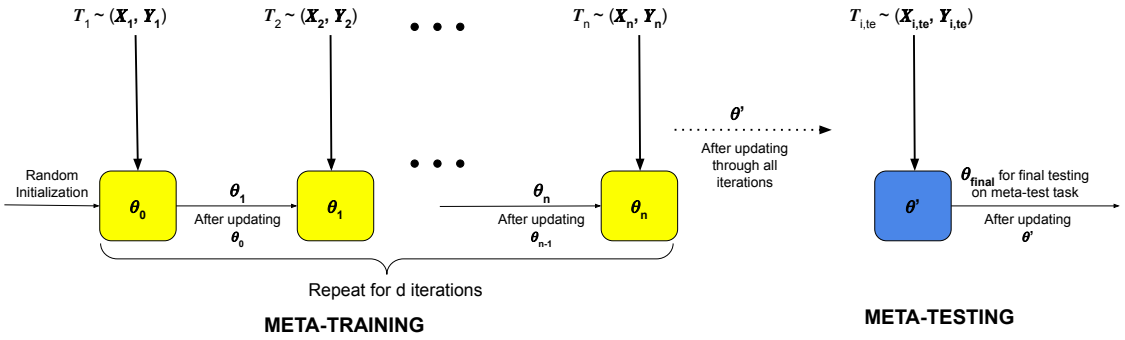


Figure 3. Pipeline for gradient-based meta-learning on skin lesion classification.

lize Group equivariant convolutions in the neural network used for Meta-DermDiagnosis to make the network invariant to any transformations to skin lesion images, which in turn improves the classification performance of the network. Now, we describe Reptile, Prototypical networks and G-convolutions in detail in following subsections.

3.1. Reptile: Gradient-based meta-learning

In supervised learning, the common practice is to learn from a set of labeled examples, while meta-learning learns from a set of labeled tasks, each represented as a labeled training set and a labeled testing set. The learning algorithm trains for a representation that can be quickly adapted to a new task, via a few gradient steps. The meta-learner seeks to find an initialization that is not only useful for adapting to various problems, but also can be adapted quickly in a small number of steps and efficiently using only a few annotated examples.

Figure 3 depicts the gradient-based meta-learning pipeline used for disease classification from skin lesion images. Let T_{tr} and T_{te} be the set of train and test tasks respectively, and D_{tr}, $train(D_{te})$, $test(D_{te})$ be the meta-train dataset, training portion of meta-test dataset used for fine-tuning, and the portion of the meta-test dataset used for testing respectively. $T_i \in T_{tr}, T_{i,te} \in T_{te}$ are the ith meta-train and meta-test tasks respectively, $(\mathbf{X}_i, \mathbf{Y}_i)$ are mini-batches sampled from D_{tr}, and $(\mathbf{X}_{i,te}, \mathbf{Y}_{i,te})$ are mini-batches sampled from $train(D_{te})$. Here, d is the number of times each task is observed during meta-training, $train(T)$ and $test(T)$ are the training and testing sets for a particular task T. The main idea is that instead of hand-designing a learning algorithm for the task of interest, we aim to learn a good network initialization (i.e., the parameters of a network) so that the classifiers for novel classes can be learned with a limited number of labeled examples and a small number of gradient update steps. The meta-learner learns an initialization from the tasks in the training set T_{tr}, with the loss being measured on the set of test tasks T_{te}. The optimization problem can be formulated using the below objective function:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{T \sim p(T)}[L_{test(T)}(\boldsymbol{\theta})] \tag{1}$$

where $L_{test(T)}(\boldsymbol{\theta})$ is the loss obtained for the test dataset of task T i.e. $test(T)$ using parameters $\boldsymbol{\theta}$.

To demonstrate the effectiveness of meta-learning on few-shot tasks, we take classes with the minimum amount of data (i.e. tail classes of the distribution) for meta-testing. Each task T_i is a classification task which is sampled from the distribution $p(T)$ over the task space. We use Reptile [25] as the meta-learning algorithm for our analysis. It is a first-order gradient-based meta-learning algorithm which learns an initialization for the parameters of a neural network model, such that when we optimize these parameters at test time, learning is fast i.e., the model generalizes using a small number of examples from the test task.

Algorithm 1 Reptile [25]

1: Initialize $\boldsymbol{\theta}$, the vector of initial parameters
2: **for** $iteration = 1, 2, \ldots$ **do**
3: Sample task T, corresponding to loss L_T on weight vectors $\tilde{\boldsymbol{\theta}}$
4: Compute $\tilde{\boldsymbol{\theta}} = U_T^k(\boldsymbol{\theta})$, denoting k SGD or Adam steps
5: Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where ϵ is the stepsize parameter
6: **end for**

In Reptile algorithm, $U_T^k(\boldsymbol{\theta})$ is the operator (e.g. corresponding to Adam optimizer or SGD) that updates $\boldsymbol{\theta}$ using k mini-batches on data sampled from T.

3.2. Prototypical Networks [31]

Prototypical network is a distance metric based meta-learning technique which computes a prototype vector as the representation of each class, and this vector is the mean vector of the embedded support instances belonging to its

class. A subset of N classes is randomly selected to formulate one training task. For each training task, a support set $S = (\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ and a query set $Q = (\mathbf{x}_{n+1}, y_{n+1}), ..., (\mathbf{x}_{n+m}, y_{n+m})$ are created by sampling examples from the selected classes, where $\mathbf{x}_j$ are inputs and y_j are the corresponding labels. Here, n and m denote the number of examples in the support (S) and query (Q) sets respectively. Prototypical Networks compute representations of the inputs $\mathbf{x}$ using an embedding function g parameterized with $\boldsymbol{\theta}$: $\mathbf{z} = g(\mathbf{x}, \boldsymbol{\theta})$. Each class c is represented in the embedding space by a prototype vector $\mathbf{m}_c$ which is computed as the mean vector of the embedded inputs for all the examples S_c of the corresponding class c as follows:

$$\mathbf{m}_c = \frac{1}{|S_c|} \sum_{(\mathbf{x}_j, y_j) \in S_c} g(\mathbf{x}_j, \boldsymbol{\theta}) \tag{2}$$

The distribution over predicted labels y for a query sample $\mathbf{x}$ is computed using softmax over negative distances to the prototypes in the embedding space using a distance function d as follows:

$$p(y = c|\mathbf{x}, \mathbf{m}_c) = \frac{\exp\left(-d(\mathbf{z}, \mathbf{m}_c)\right)}{\sum_{c'} \exp\left(-d(\mathbf{z}, \mathbf{m}_{c'})\right)} \tag{3}$$

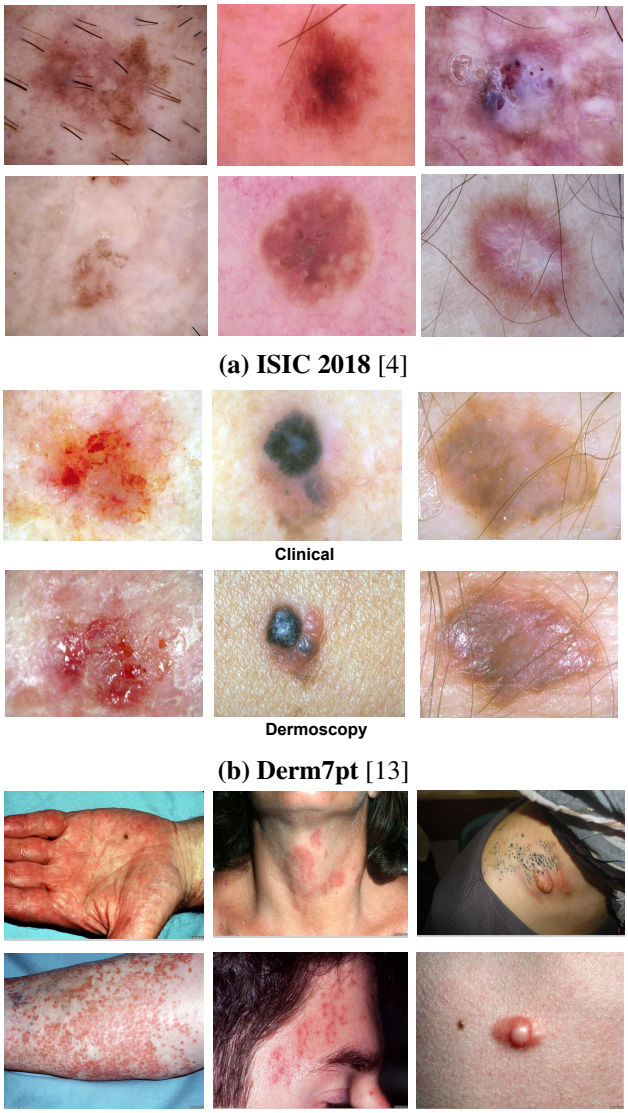
Here, $\mathbf{z} = g(\mathbf{x}, \boldsymbol{\theta})$. Parameters $\boldsymbol{\theta}$ are updated so as to improve the likelihood computed on the query set:

$$\sum_{(\mathbf{x}_j, y_j) \in S_c} \log p(y = y_j|\mathbf{x}_j, \mathbf{m}_c) \tag{4}$$

which is computed using (3) with the estimated prototypes.

3.3. Group Equivariant Convolutions

In addition, we make use of Group equivariant convolutions [5] (G-convolutions) in our neural network architecture in place of the normal spatial convolution filters. Convolutional neural networks have an important property of translational weight sharing which imparts translation equivariance. This means that shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature maps. However, there are many scenarios in which the input image consists of patterns which maintain their identity under other transformations like rotation and reflection such as images of lesions or tumors. To make the conventional CNNs equivariant to other transformations, a large amount of annotated data is required, which is difficult to obtain especially for data-limited domains such as healthcare. Thus, there is a need for data-efficient modifications such as G-convolutions through which CNNs can be generalized to other kinds of transformations. By exploiting symmetries, Group Equivariant CNNs have achieved state-of-the-art results on a variety of datasets like rotated MNIST [19] and CIFAR10 [17]. Due to the rotations and reflections present in the skin lesion data, G-convolutions enhance the performance significantly.



(a) ISIC 2018 [4]

(b) Derm7pt [13]

(c) SD-198 [32]

Figure 4. Figure showing some sample images from skin lesion datasets.

4. Datasets and Evaluation

We tested our proposed meta-learning based approach Meta-DermDiagnosis on three publicly available datasets:

ISIC 2018 Skin Lesion dataset [4] consists of $10,015$ dermoscopic images which have been labelled by expert pathologists into one of the seven categories of skin lesions. Out of the total images $7,515$ belong to the train set, while

Table 1. Performance comparison of AUC (in %) and Accuracy (in %) on the **ISIC 2018** skin lesion dataset for 2-way classification tasks.

		Pre-trained		**Reptile**		**Prototypical Networks**	
		Avg. AUC	Avg. Accuracy	Avg. AUC	Avg. Accuracy	Avg. AUC	Avg. Accuracy
w/o G-Conv	**1-shot**	59.7	54.8	60.3	58.0	**61.6**	**59.3**
	3-shot	67.8	65.2	**73.1**	**73.4**	70.2	67.9
	5-shot	72.0	71.5	**79.6**	**76.2**	75.4	73.0
w/ G-Conv	**1-shot**	61.3	62.6	**68.1**	64.3	65.7	**64.5**
	3-shot	72.8	69.3	**81.2**	**76.7**	75.8	73.5
	5-shot	79.1	79.4	**86.8**	**82.1**	82.9	79.7

the remaining $2,500$ belong to the test set according to the standard train/test split. In most cases, albeit not always, the target lesion is in the center of the image. In practice, the task of the clinician is not only to differentiate between malignant and benign lesions, but also to make specific diagnoses because different malignant lesions, for example melanoma and basal cell carcinoma, may be treated in a different way and timeframe. With the exception of vascular lesions, which are pigmented by haemoglobin and not by melanin, all lesions have variants that are completely devoid of pigment (for example amelanotic melanoma). For experimentation, we resize the images from 600×450 pixels to 224×224 pixels, and choose four and three classes in the meta-train and meta-test sets respectively for creating few-shot classification tasks. Figure 4(a) shows sample images from the dataset.

Derm7pt [13] is a dataset that includes over 2000 clinical and dermoscopy color images belonging to 20 classes, along with corresponding structured metadata which is tailored for training and evaluating computer aided diagnosis (CAD) systems. It provides image-based prediction based on a 7-point skin lesion malignancy checklist. The original image size is 768×512 pixels, and these are resized to 224×224 pixels for conducting our experiments. We have used the standard dataset train/test split in our experiments. 2 classes have been removed from our experiments: 'miscellaneous' (since this stands for random skin diseases not in the list of specified diseases) and 'melanoma' (since it has just a single example so there is no train/test split for this category). Out of 18 lesion categories, 13 classes have been used for training and the remaining classes are used for testing. The classes with the minimum amount of data have been put in the test set to model the generalization to rare skin diseases. Few example skin lesion images are shown in Figure 4(b). The first row shows sample clinical images, while the second row shows the corresponding dermoscopy images.

SD-198 [32] dataset contains 198 fine-grained skin disease categories from different types of eczema, acne and various cancerous conditions. It consists of clinical skin disease images submitted by patients or dermatologists. There are $6,584$ images in total. The images vary in color, exposure, illumination and scale, and include a wide range of patients with different ages, gender, locations of disease, colors of skin and different stages of the disease. We use the standard $50-50$ train/test split provided by SD-198, which has $3,292$ training and $3,292$ testing images. The images are collected by digital cameras or mobile phones. The original image size is 1640×1130 pixels, and they are resized to 224×224 pixels. In our experiments, we use 20 classes for training and the 70 classes which contain less than 20 images per class for testing. These 70 classes signify rare diseases for which not enough images are available. Figure 4(c) shows representative images from the dataset.

We apply augmentations to all the above datasets by applying random transformations like rotation (-30 degree to $+30$ degree), scaling (-20% to $+20\%$), and horizontal flipping. An obvious problem with datasets is heavy class imbalance (i.e., long-tailed class distribution) as shown with blue color in Figure 1. To mitigate this issue, we choose the classes with very few samples (i.e., tail classes of the distribution which refer to rare / new diseases) to create meta-test tasks.

5. Experiments and Results

5.1. Implementation Details

We use a 6-layer CNN for all our experiments involving Reptile, pre-trained networks, and Prototypical networks. Each convolution layer consists of 32 filters of size 3×3, and is followed by a 2×2 max pooling layer, batch normalization, and ReLU activation. Experiments with deeper networks like DenseNet-121 caused the models to over-fit during meta-training and few-shot fine-tuning. We performed grid search to determine the optimal learning rate and number of iterations required for meta-training and fine-tuning for all the experiments conducted, since these hyper-parameters have a considerable impact on the model

Table 2. Performance comparison of AUC (in %) and Accuracy (in %) on the **Derm7pt** skin lesion dataset for 2-way classification tasks.

		Pre-trained		**Reptile**		**Prototypical Networks**	
		Avg. AUC	Avg. Accuracy	Avg. AUC	Avg. Accuracy	Avg. AUC	Avg. Accuracy
w/o G-Conv	**1-shot**	56.9	58.4	59.7	60.2	**60.6**	**62.5**
	3-shot	62.1	60.7	64.1	**65.7**	**65.8**	63.9
	5-shot	66.6	64.9	**71.4**	**70.5**	68.2	66.7
w/ G-Conv	**1-shot**	60.8	59.5	62.1	61.8	**63.7**	**64.1**
	3-shot	62.6	62.3	**68.7**	**69.9**	65.3	66.8
	5-shot	69.8	65.2	**77.2**	**76.9**	72.8	69.5

performance. The input to the network is an image of size 224×224 pixels, and the batch size is set to 5. The standard train/test splits are used for all the 3 datasets - ISIC 2018, Derm7pt, SD-198. In addition, to incorporate G-convolutions into the network architecture, we simply replace the traditional spatial convolution layers with the G-convolution layers, with the rest of the parameters being exactly the same. The specific implementation details for the various approaches used are as follows:-

Reptile: We created binary classification tasks for meta-training and meta-testing stages for each of the 3 datasets due to the fewer number of total classes in the datasets. For the SD-198 dataset which contains 198 classes of which 90 have been used for experimentation (i.e., 20 train classes and 70 test classes), we have additionally experimented with 4-way classification tasks. We query 15 images from the meta-train dataset for each of the classes in a task during the meta-training stage. During meta-testing, k images are sampled from the training split of each class involved in the meta-test task. The value of k in our experiments is 1, 3, and 5 indicating 1-shot, 3-shot, and 5-shot respectively. These images are used for fine-tuning the model obtained as a result of meta-training. The final inference is performed on the entire testing split of the classes in the meta-test task to compute the accuracy and AUC values.

Prototypical Network: We trained prototypical networks using Euclidean distance, with the training episodes containing 4, 13, and 20 classes for the ISIC, Derm7pt, and the SD-198 datasets respectively. The train-shot is set to 15, which means that during the n-way training, 15 images per class are randomly sampled per episode from the n classes, and subsequently the model is trained on these images. The models were trained via SGD with Adam [14] optimizer. At test time on the meta-test skin lesion datasets, we created 2-way classification tasks. We also experimented with 4-way classification on the SD-198 meta-test dataset. Eventually, analysis is conducted on the average accuracy and AUC values for the test tasks. In each episode during the meta-test stage, n classes are randomly selected from the meta-test dataset to construct an n-way classification task and 1, 3, and 5 support points are selected per class within each task corresponding to 1-shot, 3-shot, and 5-shot classification. The final results are computed on the complete test dataset for a particular task.

Baseline: Pre-trained Network: The pre-trained network involves training a neural network on the entire training dataset of all the train classes, and subsequently fine-tuning and evaluating on test classes. The fine-tuning stage involves creating classification tasks, sampling k (1, 3, 5 for 1-shot, 3-shot, 5-shot respectively) images each from the train splits of the classes in each task, fine-tuning the model on the task, and finally evaluating the task on the test split of the task. We create 2-way classification tasks for all 3 datasets, and 4-way classification tasks for the SD-198 dataset. The average accuracy and AUC value is used for the performance analysis.

5.2. Experimental Analysis

Comparative results for the various techniques on ISIC, Derm7pt, and SD-198 datasets are summarized in Tables 1, 2, 3 respectively. In some 1-shot learning cases like for the ISIC and Derm7pt datasets, the performance of prototypical networks is slightly better than Reptile. However, for most cases Reptile outperforms prototypical networks significantly. The reason for prototypical network's better performance over Reptile for 1-shot learning is that in some cases, the neural network is not able to fine-tune effectively on 1 single data sample and is thus unable to generalize on the test data. For slightly higher number of samples, Reptile outdoes prototypical networks since the availability of more data samples allows the neural network to be fine-tuned effectively without over-fitting. In order to further validate the idea that fine-tuning allows better generalization on novel classes, we tried fine-tuning the prototypical network during the meta-test stage before computing the prototype vectors. It was found that the per-

Table 3. Performance comparison of AUC (in %) and Accuracy (in %) on the **SD-198** skin lesion dataset for 2-way classification tasks.

		Pre-trained		**Reptile**		**Prototypical Networks**	
		Avg. AUC	Avg. Accuracy	Avg. AUC	Avg. Accuracy	Avg. AUC	Avg. Accuracy
w/o G-Conv	**1-shot**	56.4	55.7	**64.1**	**63.0**	59.4	59.8
	3-shot	65.3	60.7	**77.4**	**72.9**	70.6	66.6
	5-shot	77.9	73.6	**84.6**	**80.4**	80.7	78.3
w/ G-Conv 2-way	**1-shot**	57.4	56.9	**68.6**	**65.3**	62.9	64.5
	3-shot	70.2	69.1	**79.1**	**75.8**	74.5	72.1
	5-shot	84.2	76.5	**89.5**	**83.7**	85.6	80.2

formance of prototypical networks improves as a result of fine-tuning, while still remaining lesser compared to Reptile. This is due to the fact that prototypical networks rely on a simple distance measure like Euclidean distance for classification which does not perform very well for complex fine-grained skin lesion images where the intra-class variability is higher compared to the inter-class variability. In such cases, a neural network based classification performs much better which is observed through our experiments. We observe empirically that 5-shot meta-learning performance surpasses 3-shot performance, which in turn exceeds the performance for 1-shot meta-learning. This trend is consistent for the baseline pre-training technique as well as the meta-learning techniques across all datasets, with both spatial and G-convolutions. The reasoning behind this observation is intuitive since more test data for fine-tuning and creating the prototype vectors leads to better model adaptation and hence, higher performance on the test classes. For instance, with G-Conv., Reptile achieves 1-shot AUC of 68.1%, 3-shot AUC of 81.2% and 5-shot AUC of 86.8% on ISIC 2018 dataset. Similarly for Derm7pt, AUC values using prototypical networks are 63.7% (1-shot), 65.3% (3-shot) and 72.8% (5-shot).

In addition to binary classification tasks, we have also experimented with creating 4-way classification tasks for the SD-198 dataset for which 4-way tasks are feasible due to the sizeable number of categories present. The 4-way setting validates all trends established by experiments so far. Reptile outperforms prototypical network which in turn performs better than the pre-training baseline. 4-way 5-shot accuracy values 65.7% (Reptile), 55.9% (Prototypical Networks), and 49.8% (Pre-training) illustrate the trend. Similarly, 4-way 3-shot accuracy values are 57.1% (Reptile), 50.4% (Prototypical Networks), and 42.6% (Pre-training). A similar trend can be seen for 4-way 1-shot and 3-shot accuracy values for the 3 different approaches.

Next, we incorporate G-convolutions into the meta-learning network architecture which enhances the network's performance on the skin lesion datasets, as can be seen in Tables 1, 2, 3 for ISIC 2018, Derm7pt and SD-198 skin lesion datasets respectively. Hence, this validates our hypothesis that G-convolutions are effective in skin lesion images as they make the neural network equivariant to image transformations.

5.3. Comparison with state-of-the-art

To the best of our knowledge, there have been no works so far that have explored few-shot learning for skin lesion datasets Derm7pt and SD-198. Thus, we provide the first work which extends meta-learning to dermoscopic and clinical skin lesion images present in the Derm7pt and SD-198 datasets. Our proposed Meta-DermDiagnosis model is able to perform better than the DAML [20] model on the ISIC 2018 dataset. We make use of the Reptile algorithm along with G-convolutions and a 6-layer CNN, while Li *et al*. propose their own difficulty-aware meta-learning (DAML) method that uses a 4-layer CNN. The meta-learning setting is similar for DAML and our approach. We outperform DAML for all 1, 3, 5 shot learning. Meta-DermDiagnosis gives a 5-shot AUC of 86.8%, while DAML gives 83.3%. The reason for this increase in performance is the more expressive 6-layer CNN, the use of Reptile algorithm and the deployment of G-convolutions which allow the network to achieve invariance to the different transformations present in skin lesion images with much fewer images.

6. Conclusion and Future Work

We propose the use of meta-learning techniques together with G-convolutions for skin disease identification and, quick and efficient model adaptation for extremely low-data scenarios. We show how our methodology outperforms the conventional transfer learning or fine-tuning approach for the data-scarce settings. We believe that further research in this direction should focus on extending meta-learning for other medical imaging datasets like X-rays, CT-Scans and MRIs which can play a vital role in diagnosing and detecting rare and new diseases like the recent COVID-19 which have limited available patient data.

References

[1] M. Z. Alom, C. Yakopcic, T. M. Taha, and V. K. Asari. Nuclei segmentation with recurrent residual convolutional neural networks based u-net (r2u-net). In *NAECON 2018 - IEEE National Aerospace and Electronics Conference*, pages 228–233, July 2018.

[2] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949, March 2016.

[3] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *ArXiv*, abs/1904.00625, 2019.

[4] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[5] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.

[6] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *ArXiv*, abs/1804.10916, 2018.

[7] Mohammad Eslami, Solale Tabarestani, Shadi Albarqouni, Ehsan Adeli, Nassir Navab, and Malek Adjouadi. Image to images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography. *ArXiv*, abs/1906.10089, 2019.

[8] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, jan 2017.

[9] Nils Gessert, Thilo Sentker, Frederic Madesta, Rüdiger Schmitz, Helge Kniep, Ivo M. Baltruschat, René Werner, and Alexander Schlaefer. Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. *CoRR*, abs/1905.02793, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[11] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. *CoRR*, abs/1905.05901, 2019.

[12] Cheng-Bin Jin, Hakil Kim, Mingjie Liu, In Ho Han, Jae Il Lee, Jung Hwan Lee, Seongsu Joo, Eunsik Park, Young Saem Ahn, and Xuenan Cui. Dc2anet: Generating lumbar spine mr images from ct scan data based on semi-supervised learning. 2019.

[13] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Gregory R. Koch. Siamese neural networks for one-shot image recognition. 2015.

[16] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *NeurIPS*, 2018.

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[18] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[19] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480, 2007.

[20] Xiaomeng Li, Lequan Yu, Chi-Wing Fu, and Pheng-Ann Heng. Difficulty-aware meta-learning for rare disease diagnosis. *arXiv preprint arXiv:1907.00354*, 2019.

[21] Yi Li and Wei Ping. Cancer metastasis detection with neural conditional random field. *CoRR*, abs/1806.07064, 2018.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015.

[23] Luke Metz, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-dickstein. Meta-learning update rules for unsupervised representation learning. 2019.

[24] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1928–1937. JMLR.org, 2016.

[25] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[26] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018.

[27] Viraj Prabhu, Anitha Kannan, Murali Ravuri, Manish Chaplain, David Sontag, and Xavier Amatriain. Few-shot learning for dermatological disease diagnosis. In Finale Doshi-Velez,

Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 532–552, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[29] Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, and Alexandr A. Kalinin. Deep convolutional neural networks for breast cancer histology image analysis. In Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny, editors, *Image Analysis and Recognition*, pages 737–744, Cham, 2018. Springer International Publishing.

[30] Taibou Birgui Sekou, Moncef Hidane, Julien Olivier, and Hubert Cardot. From patch to image segmentation using fully convolutional networks - application to retinal images. *ArXiv*, abs/1904.03892, 2019.

[31] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[32] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222. Springer, 2016.

[33] Ahmed Taha, Pechin Lo, Junning Li, and Tao Zhao. Kid-net: Convolution networks for kidney vessels segmentation from ct-volumes. In *MICCAI*, 2018.

[34] Youbao Tang, Yuxing Tang, Jing Xiao, and Ronald M. Summers. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation. In *MIDL*, 2019.

[35] Joaquin Vanschoren. Meta-learning: A survey. *CoRR*, abs/1810.03548, 2018.

[36] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3637–3645, Red Hook, NY, USA, 2016. Curran Associates Inc.

[37] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J. Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In *Neural Information Processing Systems*, 2019.

[38] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. T. K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, pages 1–1, 2019.

[39] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.