META-TRANSFER DERM-DIAGNOSIS: EXPLORING FEW-SHOT LEARNING AND TRANSFER LEARNING FOR SKIN DISEASE CLASSIFICATION IN LONG-TAIL DISTRIBUTION

Zeynep Özdemir
Ankara University
Computer Engineering Department
zynpozdemir@ankara.edu.tr

Hacer Yalim Keles
Hacettepe University
Computer Engineering Department
hacerkeles@cs.hacettepe.edu.tr

Ömer Özgür Tanrıöver
Ankara University
Computer Engineering Department
tanriover@ankara.edu.tr

ABSTRACT

Addressing the challenges of rare diseases is difficult, especially with the limited number of reference images and a small patient population. This is more evident in rare skin diseases, where we encounter long-tailed data distributions that make it difficult to develop unbiased and broadly effective models. The diverse ways in which image datasets are gathered and their distinct purposes also add to these challenges. Our study conducts a detailed examination of the benefits and drawbacks of episodic and conventional training methodologies, adopting a few-shot learning approach alongside transfer learning. We evaluated our models using the ISIC2018, Derm7pt, and SD-198 datasets. With minimal labeled examples, our models showed substantial information gains and better performance compared to previously trained models. Our research emphasizes the improved ability to represent features in DenseNet121 and MobileNetV2 models, achieved by using pre-trained models on ImageNet to increase similarities within classes. Moreover, our experiments, ranging from 2-way to 5-way classifications with up to 10 examples, showed a growing success rate for traditional transfer learning methods as the number of examples increased. The addition of data augmentation techniques significantly improved our transfer learning based model performance, leading to higher performances than existing methods, especially in the SD-198 and ISIC2018 datasets. All source code related to this work will be made publicly available soon at the provided URL.

Keywords Few-shot learning · Long-tail distribution · Medical image classification · Skin disease classification · Transfer learning

1 Introduction

Over the past decade, the field of medical image analysis has witnessed remarkable advancements, primarily driven by the development of deep convolutional neural networks and the availability of extensive labeled image datasets. These advancements have notably impacted various tasks, including organ segmentation [1, 2], tumor segmentation [3, 4], and disease detection [5, 6]. Although abundant data exists for common diseases, a significant gap persists in data availability for the over 6,000 known rare diseases, affecting approximately 7% of the global population [7]. The diagnosis of these rare diseases, including some skin conditions, presents unique challenges, particularly due to the limited number of clinical examples available for training deep learning models. The automatic classification of skin lesions exemplifies these challenges, as it is complicated by the long-tailed distribution of skin disease datasets, the subtle variations in lesion appearances, and the overall scarcity of sufficient image data [8].

Various studies have been conducted to address the problem of skin disease classification using deep learning approaches. Recent advancements in this field are mainly in three categories: methods based on transfer learning [12, 13], those relying on few-shot learning [8, 14–19], and approaches using cross-domain few-shot learning [20]. The state of the art models in this domain, such as Meta-DermDiagnosis, MetaMed, and PCN models [8, 14, 18], are designed to extract and learn high-level, domain-specific features during their training process.

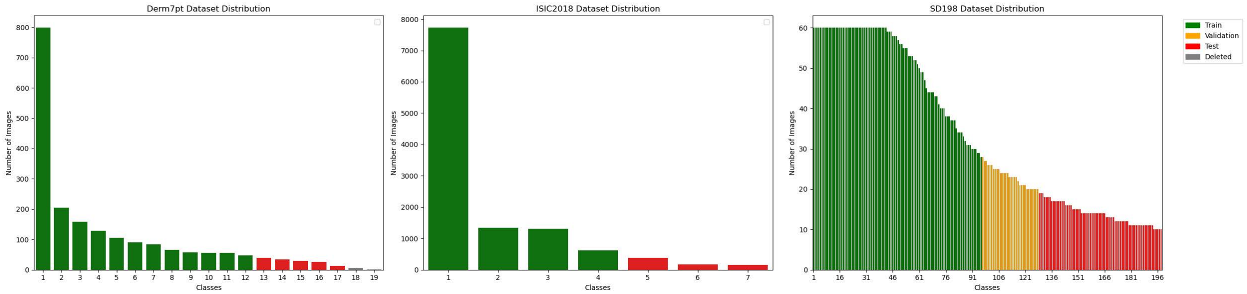

Figure 1: The figure illustrates the distributions of the datasets used in this study, namely SD-198 [9], Derm7pt [10] and ISIC2018 [11] . The observed data distribution exhibits a long-tailed nature, and an examination reveals that some classes have very few instances. The base classes (common diseases) used during the training process include the train (green) and validation (yellow) classes. Novel classes (new/rare diseases) are indicated with the test label (red). In the Derm7Pt dataset, classes with a significantly low number of examples are displayed as deleted (grey) and deemed unusable.

In their study on the ISIC2018 dataset, Li et al. (2020) introduced the Difficulty Aware Meta-Learning (DAML) model, recognizing that randomly selected tasks vary in difficulty [15]. They proposed optimizing the meta-optimization process by dynamically assessing the significance of challenging tasks. Similarly, Mahajan et al. (2020) developed the Meta-DermDiagnosis model, experimenting with the SD-198, Derm7pt, and ISIC2018 datasets [8]. This model innovatively replaced traditional convolutional layers in Prototypical Networks and Reptile with Group Equivariant Convolutions, enhancing resistance to data transformations like symmetry and rotation, thus yielding robust features. However, this method received critique for its reliance on datasets with symmetric orientation. On a related note, Singh et al. (2021) highlighted the efficacy of MixUp, CutOut, and CutMix as data augmentation techniques in the medical field when integrated with the MAML algorithm in their MetaMed model, particularly on the ISIC2018 dataset [18]. Both MetaMed and Meta-DermDiagnosis models aim to enhance feature representation by broadening data augmentation and diversification. Further, Prabhu et al. (2019) proposed the Prototypical Clustering Networks (PCN) model, designed to effectively handle intra-class variability in few-shot learning scenarios [14].

The studies previously discussed have employed episodic learning as a means to acquire knowledge that can be transferred to new classes. This approach involves the arrangement of learning problems into episodes, each composed of small training and validation subsets, designed to simulate the scenarios encountered during evaluation. However, [21] criticized this method, arguing that the constraints of episodic learning are unnecessary and that using training groups in this manner is data-inefficient. Similarly, [22] argued that in tasks involving rare classes, the effectiveness of rapid adaptation is more attributable to the quality of the learned representation than to the few-shot learning algorithm itself. They incorporated self-supervised learning in their model to enhance this representation. In a similar context, [23] proposed the concept of Meta-transfer learning, which is aimed at refining the learned representation. They conducted various experiments to compare the effectiveness of fine-tuning pre-trained Deep Neural Network (DNN) models with and without episodic learning. Their findings underscore the significance of integrating transfer learning with few-shot learning methodologies.

The existing literature consistently highlights the ongoing challenge of effectively managing the issue of long-tailed data distribution in skin disease studies [8, 24, 25]. Although current methods show promise in specific dataset contexts, their ability to generalize broadly remains limited. Most research in this field has been directed towards developing dataset-specific solutions, addressing the unique challenges and sensitivities arising from variations in class counts, data formats, and other critical characteristics.

In our study, we aim to conduct an in-depth analysis of the benefits and drawbacks of integrating few-shot learning, episodic learning, and transfer learning. Our main focus is on developing a foundational framework specifically designed to utilize the inherent properties of long-tailed skin datasets. While acknowledging the potential benefits of performance-enhancing methods, our base approach distinctly diverges from such techniques.

In this context, we implemented four distinct training methods, each evaluated using consistent metrics for newly introduced classes. Our first method, Few-Shot Episodic Transfer Learning (FETL), involves adapting the model for the dataset through fine-tuning weights that were initially pre-trained on ImageNet, coupled with episodic few-shot learning. The second strategy, Few-Shot Episodic Learning (FEL), diverges from the first by employing episodic learning on models without any pre-training. The third approach, Deep Transfer Learning (DTL), focuses on fine-tuning models using deep neural networks (DNN) with pre-trained ImageNet weights, but notably omits episodic learning. Lastly, our fourth method, Deep Learning (DL), acts as a baseline, utilizing pre-trained ImageNet weights without additional adjustments or fine-tuning. Following the establishment of our approach, we aimed to enhance performance by incorporating well-known data augmentation techniques such as CutMix, MixUp, and ResizeMix into the DTL model.

To evaluate the effectiveness of our proposed methodologies, we carried out extensive testing across three benchmark skin disease datasets: SD-198, Derm7pt, and ISIC2018. This comprehensive analysis provided us with critical insights into the most effective strategies for tackling the challenge of long-tailed data distribution in skin diseases. The key contributions of our study are summarized as follows:

- We introduce a novel methodology designed to evaluate various model training approaches, using a consistent benchmark test set specifically made for long-tail distributions in rare skin diseases. This evaluation is conducted through episodic testing. To the best of our knowledge, this is the first time such a thorough methodological analysis has been conducted in this domain.
- By comparing episodic and traditional training methods, our findings indicate that traditional training becomes increasingly beneficial as the number of shots (training examples) grows.
- We demonstrated that combining transfer learning with few-shot learning significantly enhances both the learned representation and the testing performance in the context of rare skin diseases. Using our proposed model based on transfer learning, along with augmentation techniques like MixUp, CutMix, and ResizeMix, has helped us surpass state-of-the-art results in some settings on the SD198 and ISIC2018 datasets.

2 Related Works

2.1 Transfer Learning for Medical Image Domain

In deep learning, a powerful transfer learning method involves adapting a pre-trained model for a new task, commonly referred to as fine-tuning (FT). Models that have been pre-trained on extensive datasets have demonstrated superior generalization performance compared to models initialized randomly [26]. Various techniques are employed to facilitate the transfer of knowledge between different source-target domains [27–30].

Transfer learning (TL) is frequently applied in addressing the classification problem of skin diseases. For instance, [31] aimed to improve performance on the ISIC2018 dataset by using ISIC2016 and ISIC2017 as source datasets and EfficientNetB0/B1, SeReNext50 as backboned models. [32] performed transfer from the ImageNet dataset to the ISIC2018 dataset using ResNet50. Similarly, [33] employed ImageNet, ResNet, and DenseNet architectures for transfer to the ISIC2017 dataset. [34], with a similar objective, experimented with the VGG16 architecture. A common inference drawn from these studies is that the use of transfer learning, particularly when ImageNet or various dermatological datasets are used as source datasets and supported by architectures like ResNet or DenseNet, contributes significantly to addressing this problem. A more comprehensive literature discussion is provided in [35].

In addition to traditional learning methods, transfer learning has also been explored in the Few-Shot Learning (FSL) domain for the classification problem of skin diseases. For instance, the MetaMed study [18], working with FSL algorithms on the ISIC2018 dataset, compared and interpreted the results of models trained through TL with their proposed method. In this process, they employed a shallow-layered architecture in the TL model and conducted training using only base classes. In contrast, as indicated in the literature, we adapted the MobileNet and DenseNet architectures to the Few-Shot Learning domain by utilizing pre-trained weights from ImageNet. We discussed detailed comparison results in Section 5.3.

2.2 Few-Shot Learning in Computer Vision

Few-shot learning (FSL) aims to recognize novel classes with only a few labeled examples, leveraging a substantial number of examples from base classes. FSL algorithms can be broadly categorized into three groups: initialization-based methods, metric learning-based methods, and hallucination-based methods.

In this context, initialization-based methods take a *learning to fine-tune* approach. They aim to acquire an effective model initialization, specifically the neural network parameters. This facilitates the adaptation of classifiers with limited labeled examples through a few gradient update steps for new classes [36–38]. Another strategy involves distance metric learning methods, embracing a *learning to compare* paradigm for few-shot classification. These methods are foundational approaches utilizing encoded feature vectors and a distance measurement metric based on the nearest-neighbor principle to assign labels. For instance, Prototypical Networks [39] utilizes Euclidean distance, Matching

Networks [40] employs cosine similarity, and Relation Networks [41] utilizes its own CNN-based measurement module for this purpose [42]. Additionally, hallucination-based methods directly address data scarcity through *learning to augment*. Here, hallucination involves generating data not derived from real examples or direct observations. The generator's objective is to transfer appearance variations present in the base classes to novel classes [43, 44].

The mentioned FSL methods adopt an episodic training approach during the training of the data (base classes). However, some studies have shown that training the data by dividing it into tasks leads to inefficient use of the available data [21, 22]. Additionally, [45] has demonstrated, contrary to the prevailing notion, through experiments conducted with benchmark datasets and fundamental FSL algorithms, that an increase in task diversity, proportional to the increase in classes and data, does not lead to an improvement in success. Therefore, we can categorize FSL approaches into two groups based on their training processes: meta-learning-based and transfer-learning (TL)-based methods. Among TL methods, S2M2-R [46], Baseline [47], PT-MAP [48], and Meta-Transfer Learning [23] train on base classes using a standard classification network and fine-tune the classifier head on episodes generated from new classes. These methods aim to train a powerful feature extractor that produces transferable features for the new class. Experimental methods have demonstrated that these approaches can achieve more effective and higher performance compared to previous FSL methods, utilizing a simpler and more efficient process. Due to the superior performance of TL-based methods, we explored this approach in conjunction with Prototypical Networks as a way to predict rare skin diseases.

2.3 Few-Shot Learning for Skin Disease Classification

The imbalanced distribution of skin disease classes and factors such as limited image availability in rare diseases necessitate the application of few-shot learning methods. As a solution to this challenge, [15] proposed the Difficulty-Aware Meta-Learning (DAML) model based on Meta-Learning. This model adjusts the losses for each task, increasing the weight of challenging tasks while decreasing that of easier tasks, thereby emphasizing and increasing the importance of difficult tasks. This study, aiming to highlight more distinctive features for each class, can be categorized as initialization-based FSL. On the other hand, the model named MetaMed, evaluated in both initialization and hallucination-based FSL categories, by [18], combines advanced data augmentation techniques such as mixup, cutout, and cutmix with the reptile model during training to enhance its generalization capabilities. Working with the ISIC2018 dataset, they compared the transfer learning approach of their proposed model with other studies, reporting an average improvement of up to 3% in performance. In the metric-based branch, several methods have been proposed. [17] advocated for the superiority of the Query-Relative loss over the Cross-Entropy loss commonly used in FSL. Additionally, [49] suggested the utilization of Temperature Networks alongside Prototype Networks. They adapted specific temperatures for different categories to reduce intra-class variability and enhance inter-class dispersion. Moreover, they applied penalization based on the proximity of query examples. [8] introduced a model named Meta-DermDiagnosis, aiming to obtain invariant features after various transformations by replacing traditional convolutional layers with group-equivariant convolutions, similar to Prototypical Networks and Reptile.

In the context of transfer-learning-based algorithms, [25] proposed a model named PFEMed, suggesting a dual-encoder structure. This brings about one encoder with fixed weights pre-trained on large-scale public image classification datasets and another encoder trained on the target medical dataset. On the other hand, [24] designed a dual-branch framework and improved performance using a model employing prototypical networks and contrastive loss. To address the challenge of observing diverse subgroups within dermatological disease clusters, [19] introduced the Sub-Cluster-Aware Network (SCAN) model. SCAN utilizes a dual-branch structure to enhance feature explanation, learning both class-specific features for disease differentiation and subgroup-related features.

In this study, we integrate Prototypical Networks from Few-Shot Learning (FSL) algorithms with transfer learning techniques. This fusion involves adapting MobileNetV2 and DenseNet121 backbone architectures, coupled with pretrained ImageNet weights, specifically to tackle the challenge of rare skin disease identification. Our experimental analysis highlights the limitations of episodic training in optimally leveraging data. Central to our approach is the establishment of a foundational framework that relies solely on the inherent properties of the long tailed datasets, without the necessity for supplementary data. This includes strategies aimed at enhancing model performance, such as extensive data augmentation techniques like MixUp, CutMix, and ResizeMix. The utilization of these techniques contributes to achieving our research objectives.

3 Datasets and Evaluation

The SD-198 dataset [9] comprises 198 detailed categories of skin diseases, including eczema, acne, rosacea, and various cancer conditions. These categories contain 6584 clinical images contributed by patients and dermatologists, showcasing diverse characteristics like color, exposure, lighting, and size. The dataset captures a wide range of patients in terms of age, gender, disease location, skin color, and disease stage. Originally divided into a 50% training set and a

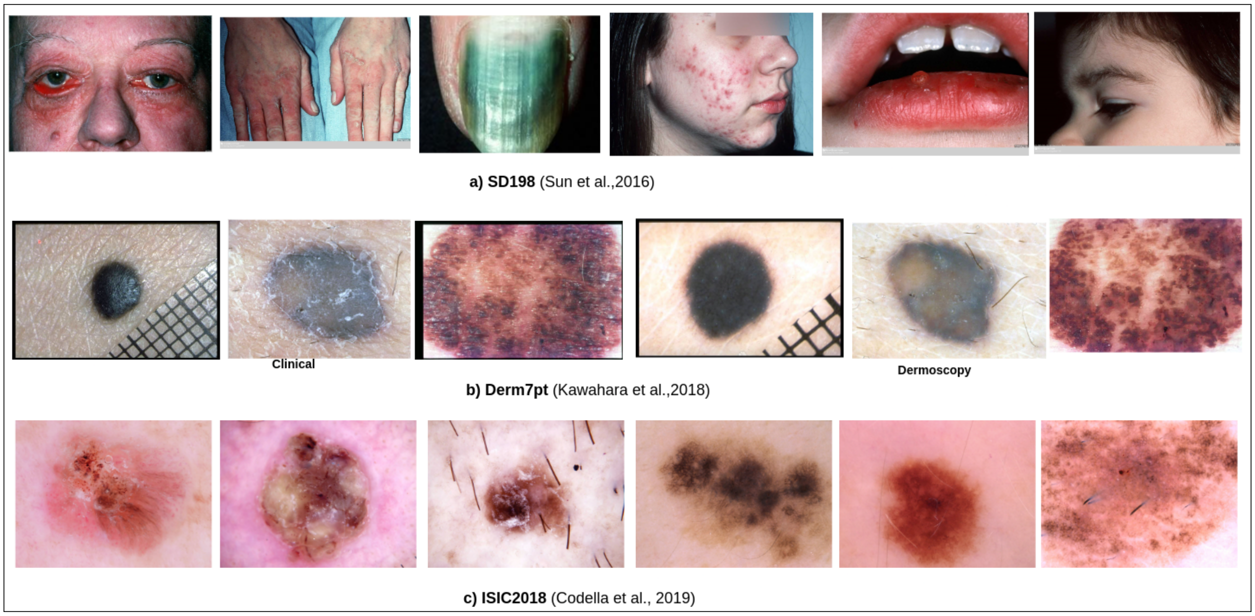
a) SD198 (Sun et al.,2016)
Clinical
Dermoscopy
b) Derm7pt (Kawahara et al.,2018)
c) ISIC2018 (Codella et al., 2019)

Figure 2: Some sample images from skin disease classification datasets

50% test set, the images are captured at 1640 × 1130 pixels using digital cameras or mobile phones. To align with the comparison criteria set by Meta-DermDiagnosis [8] and SCAN [19], we resized all data to 224 x 224 pixels. For testing, we focused on 70 classes representing rare diseases, each having fewer than 20 images, while the remaining 128 classes were used for training. The dataset distribution is visualized in Figure 1, and sample images are provided in Figure 2-a.

The Derm7pt dataset [10] comprises over 2,000 clinical and dermoscopy images grouped into 20 distinct classes. This dataset, comprising both clinical and dermoscopic images, provides predictions based on a 7-point checklist for the malignancy of skin lesions, making it suitable for training and evaluating computer-aided diagnosis (CAD) systems. The original images have dimensions of 768 × 512 pixels; however, for the purpose of our experimental studies, they have been resized to 224 × 224 pixels. To facilitate the comparison of our experimental results with the Meta-DermDiagnosis and SCAN study, we adopted similar train-test set differentiations. Within the Derm7pt dataset, two categories are excluded from our experiments: 'miscellaneous' (encompassing unspecified skin diseases) and 'melanoma' (due to its solitary instance, preventing a train-test division). Among the 18 lesion categories in this dataset, 13 classes are allocated for training, while the remaining categories are reserved for testing. The novel set consists of 5 classes, with each class containing 10 to 34 images. This distinction strategy aims to mimic the ability to generalize to infrequent skin diseases by placing classes with limited data in the test set. For visual reference, selected examples of skin lesion images can be found in Figure 2-b. The dataset distribution is also shared in Figure 1.

The ISIC 2018 Skin Lesion dataset [11] comprises 10,015 dermoscopic images that are categorized by expert pathologists into seven distinct skin lesion classes. Within this dataset, 7,515 images are allocated to the training set, while the remaining 2,500 images constitute the test set, following a standardized partition. Dermoscopic images often position the target lesion at the center. We adhered to similar scaling and data partitions as the Meta-DermDiagnosis [8] and PFEMed [25] study for our experiments. Consequently, the image resizing process transforms images from 600 × 450 pixels to 224 x 224 pixels, and a subset comprising four base classes and three novel classes is selected to form few-shot classification tasks. Refer to Figure 2-c for exemplar images drawn from the dataset. For data distribution, please see Figure 1.

4 The Methodology

This section outlines the core methodologies deployed in our study, starting with the Meta-Transfer Derm-Diagnosis Framework. This framework is pivotal for assessing four distinct model training strategies that are particularly valuable in scenarios with limited data samples. These strategies are Few-Shot Episodic Transfer Learning (FETL), Few-Shot Episodic Learning (FEL), Deep Transfer Learning (DTL), and Standard Deep Learning (DL).

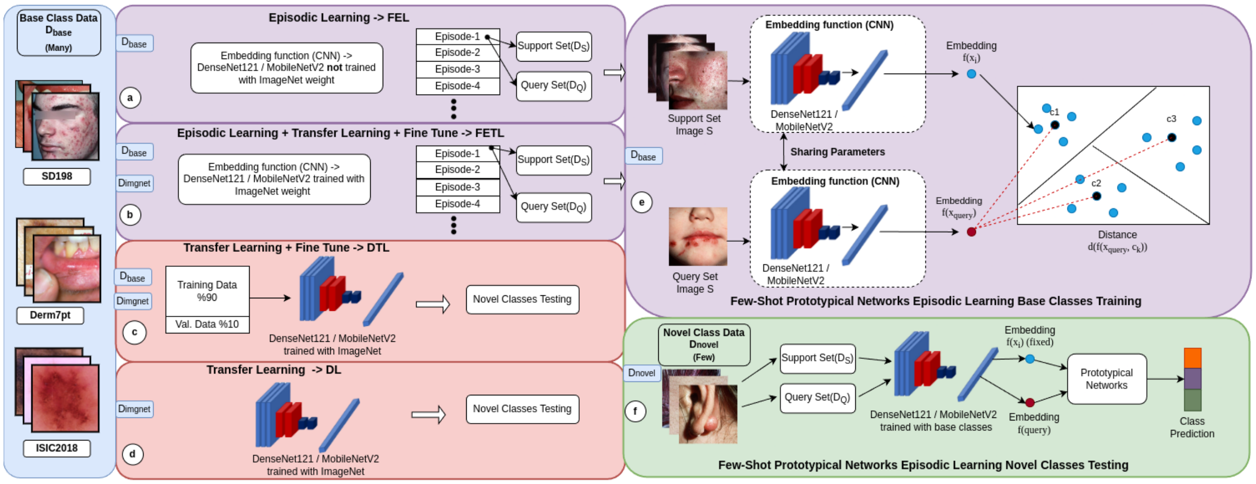
Base Class Data
D_{base}
(Many)
SD198
Derm7pt
ISIC2018
Episodic Learning -> FEL
D_{base}
a
Embedding function (CNN) -> DenseNet121 / MobileNetV2 **not** trained with ImageNet weight
Episode-1
Episode-2
Episode-3
Episode-4
Support Set(D_S)
Query Set(D_Q)
Episodic Learning + Transfer Learning + Fine Tune -> FETL
D_{base}
Dimgnet
b
Embedding function (CNN) -> DenseNet121 / MobileNetV2 trained with ImageNet weight
Episode-1
Episode-2
Episode-3
Episode-4
Support Set(D_S)
Query Set(D_Q)
Transfer Learning + Fine Tune -> DTL
D_{base}
Dimgnet
c
Training Data %90
Val. Data %10
DenseNet121 / MobileNetV2 trained with ImageNet
Novel Classes Testing
Transfer Learning -> DL
Dimgnet
d
DenseNet121 / MobileNetV2 trained with ImageNet
Novel Classes Testing
D_{base}
e
Embedding function (CNN)
Support Set Image S
DenseNet121 / MobileNetV2
Embedding $f(x_i)$
Sharing Parameters
Embedding function (CNN)
Query Set Image S
DenseNet121 / MobileNetV2
Embedding $f(x_{query})$
c1
c2
c3
Distance $d(f(x_{query}, c_k))$
Few-Shot Prototypical Networks Episodic Learning Base Classes Training
Novel Class Data
Dnovel
(Few)
Dnovel
f
Support Set(D_S)
Query Set(D_Q)
DenseNet121 / MobileNetV2 trained with base classes
Embedding $f(x_i)$ (fixed)
Embedding f(query)
Prototypical Networks
Class Prediction
Few-Shot Prototypical Networks Episodic Learning Novel Classes Testing

Figure 3: Overall framework of our pipeline: Meta-Transfer Derm-Diagnosis. a) Episodic learning is combined with DenseNet and MobileNet architectures without the use of ImageNet weights. b) ImageNet pre-trained weights are utilized along with the application of an episodic learning strategy. c) Pre-trained weights and all base class data are employed for fine-tuning with ImageNet. d) Only ImageNet weights are utilized without fine-tuning. e) Detailed diagram illustrating the use of episodic learning and prototypical networks. It is applied in the continuation of parts a and b. f) Common evaluation scheme using novel data across segments a, b, c, and d.

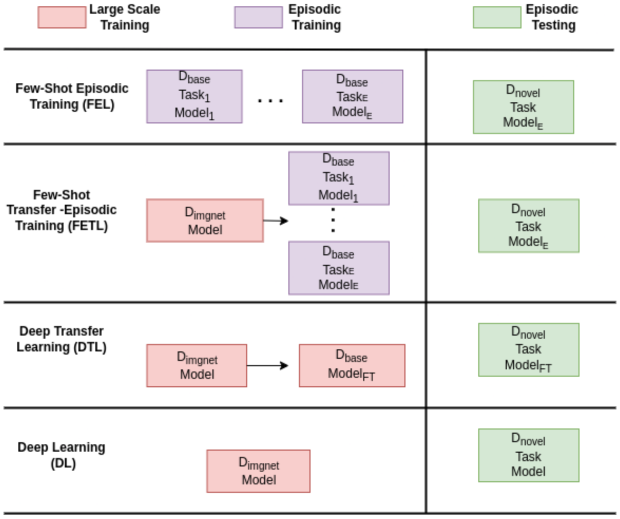
Large Scale Training
Episodic Training
Episodic Testing
Few-Shot Episodic Training (FEL)
D_{base} $Task_1$ $Model_1$
. . .
D_{base} $Task_E$ $Model_E$
D_{novel} Task $Model_E$
Few-Shot Transfer -Episodic Training (FETL)
D_{imgnet} Model
D_{base} $Task_1$ $Model_1$
D_{base} $Task_E$ $Model_E$
D_{novel} Task $Model_E$
Deep Transfer Learning (DTL)
D_{imgnet} Model
D_{base} $Model_{FT}$
D_{novel} Task $Model_{FT}$
Deep Learning (DL)
D_{imgnet} Model
D_{novel} Task Model

Figure 4: The flowchart illustrating the components and the training strategies of the FEL, FETL, DTL, and DL models within the proposed framework.

We implement a precise evaluation process, employing a combination of the selected benchmark datasets (Section 3), and a specialized testing approach. This ensures a comprehensive and fair analysis of each training method. Subsequently, we provide an overview of Prototypical Networks and Transfer Learning.

4.1 Meta-Transfer Derm-Diagnosis Framework

The proposed Meta-Transfer Derm-Diagnosis Framework is used to effectively combine few-shot training methods with transfer learning. This integration is specifically designed to improve model performance for tail classes, which often have limited data, thereby leveraging the complementary strengths of both methodologies.

Few-shot classification operates with two distinct datasets: the base dataset, denoted as D_{base}, and the novel dataset, denoted as D_{novel}. The novel dataset, D_{novel}, is utilized for the actual classification task, while the base dataset, D_{base}, helps in training the classifier by transferring essential knowledge from it. Additionally, during training, ImageNet dataset is also used and will be referred to as the domain D_{imgnet}. For clarity and coherence, let us include two definitions adapted to our framework from prior literature [42] that are related to our study's context in few-shot learning. These definitions will help in framing our approach and methodology.

Definition 1. For the training and testing phases, the novel dataset D_{novel} is split into two subsets: the support set (D_S) and the query set (D_Q). In a typical **few-shot classification** scenario, the support set D_S contains only a few samples per class, often ranging from 1 to 5. The primary goal in few-shot classification is to train a classifier, $f : X_{\text{novel}} \rightarrow Y_{\text{novel}}$, using the limited data in D_S. This classifier should then be able to accurately categorize instances in the query set D_Q. If D_S includes N distinct classes with K labeled examples per class, this scenario is defined as an N-way K-shot classification. The case with only one labeled example per class is termed one-shot classification.

Definition 2. A few-shot classification task is referred to as *cross-domain few-shot classification* when the train dataset D_{train} and the novel dataset D_{novel} are sourced from distinct domains.

To further clarify our methodology, we define the datasets used. The base dataset is defined as $D_{\text{base}} = \{(x_i, y_i); x_i \in X_{\text{base}}, y_i \in Y_{\text{base}}\}_{i=1}^{N_{\text{base}}}$, where x_i represents the feature vector of the i-th image, and y_i is its corresponding class label. The novel dataset is similarly represented as $D_{\text{novel}} = \{(\tilde{x}_j, \tilde{y}_j); \tilde{x}_j \in X_{\text{novel}}, \tilde{y}_j \in Y_{\text{novel}}\}_{j=1}^{N_{\text{novel}}}$. It is crucial to note that the class labels in D_{base} and D_{novel} are mutually exclusive, i.e., $Y_{\text{base}} \cap Y_{\text{novel}} = \emptyset$. In a similar manner, D_{imgnet} is represented as: $\{(\bar{x}_k, \bar{y}_k); \bar{x}_k \in X_{\text{imgnet}}, \bar{y}_k \in Y_{\text{imgnet}}\}_{k=1}^{N_{\text{imgnet}}}$.

To rigorously evaluate the framework we devised, detailed in Figure 3, we maintained a consistent evaluation by keeping D_{novel} fixed. Here, D_{train} denotes the datasets used during training. We formulated four distinct training methodologies, each independently designed: FETL, where $D_{\text{train}} = D_{\text{imagenet}} + D_{\text{base}}*$; FEL, where $D_{\text{train}} = D_{\text{base}}*$; DTL, where $D_{\text{train}} = D_{\text{imagenet}} + D_{\text{base}}$; and DL, where $D_{\text{train}} = D_{\text{imagenet}}$. The notation $D_{\text{base}}*$ indicates episodic training, while the others involve traditional large-scale training approaches for the corresponding datasets in that domain.

In accordance with these definitions, as is summarized in Figure 4, our FETL, FEL, and DTL models utilize D_{base} and D_{novel} classes from the same domain for training and testing as specified in Definition 1. On the other hand, the proposed DL model utilizes two separate domains for training and testing, hence it aligns with Definition 2. Therefore, three out of our four proposed analysis include adapted few-shot training and testing methodologies considering the data extracted carefully from the same long-tail distributions (FETL, FEL and DTL), while the last one (DL) corresponds to a cross-domain evaluation.

Algorithm 1 N-Way K-Shot Classification Evaluation.

```
Require: $D_{\text{train}} = \{(x_i, y_i); X_i \in \mathcal{X}_{\text{train}}, Y_i \in \mathcal{Y}_{\text{train}}\}_{i=1}^{N_{\text{train}}}$.
Require: $D_{\text{novel}} = \{(\tilde{x}_j, \tilde{y}_j); \tilde{x}_j \in \mathcal{X}_{\text{novel}}, \tilde{y}_j \in \mathcal{Y}_{\text{novel}}\}_{j=1}^{N_{\text{novel}}}$.
Require: Number of episodes $E$
for $e = 1, ..., E$ do
    Randomly select $N$ classes from $\mathcal{Y}_{\text{novel}}$.
    Randomly select $K$ samples from each class as the support set $D_{\text{S}}^{(e)}$.
    Randomly select $M$ samples from the remaining samples of $N$ classes as the query set $\{(\tilde{x}^{(e)}, \tilde{y}^{(e)})\}$.
    Record predicted labels $\hat{y}^{(e)} = f(\tilde{x}^{(e)} | D_{\text{train}}, D_{\text{S}}^{(e)})$.
    Compute accuracy $a^{(e)} = \frac{1}{M}\sum_{m=1}^{M} 1[\hat{y}^{(e)} = \tilde{y}^{(e)}]$
end for
Compute: $Avg\_Acc = \frac{1}{E}\sum_{e=1}^{E} a^{(e)}$
return $Avg\_Acc$
```

The evaluation of our classifier in N-way K-shot classification is outlined in Algorithm 1. This process includes a series of episodes, each presenting a unique classification task, allowing for a thorough assessment of the classifier's performance. In this procedure, as the initial step, we randomly select N classes from the novel set. Following this, we randomly choose K samples from each of these N classes to constitute a support set. Concurrently, we select M samples from the remaining instances in these classes to form a query set. For each episode, denoted as the e'th episode, the query set's instances and labels are represented as $\tilde{x}^{(e)}$ and $\tilde{y}^{(e)}$, respectively. A learning algorithm is then applied, utilizing the model that is pretrained with D_{train} and tuned to the support set of the e'th episode, $D_S^{(e)}$. This algorithm yields a classifier that predicts labels for the instances in the query set. To quantify the classifier's performance, we calculate the classification accuracy for each episode, referred to as $a^{(e)}$. The overall effectiveness of the learning algorithm is then determined by averaging these classification accuracies across all episodes.

4.2 Prototypical Networks

As depicted in Figure 3 (top right section), we used Prototypical Networks [39] for both episodic training and testing. It is a meta-learning approach, which is designed to represent each class through a prototype vector, based on distance metrics. This vector is an average of embedded instances in a support set, specifically linked to that class. Formally, for a set of N classes, the support set $S = \{(x_1, y_1), ..., (x_N, y_N)\}$ is constructed, where each $x_i \in \mathbb{R}^D$ is a D-dimensional feature vector, and y_i is its corresponding label in the range $\{1, ..., K\}$.

Each class prototype $c_k \in \mathbb{R}^M$ is computed as the mean vector of its associated embedded instances, using an embedding function $f_\phi : \mathbb{R}^D \to \mathbb{R}^M$ with learnable parameters ϕ. The prototype for class k is derived as:

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\phi(x_i). \tag{1}$$

The network utilizes a distance function $d : \mathbb{R}^M \times \mathbb{R}^M \to [0, +\infty)$ to calculate the probability distribution across classes for a query point x, based on the softmax of distances between the query point and class prototypes:

$$p_\phi(y = k|x) = \frac{exp(-d(f_\phi(x), c_k))}{\sum_{k'} exp(-d(f_\phi(x), c_{k'}))}. \tag{2}$$

Training involves minimizing the negative log likelihood $J(\phi) = -logp_\phi(y = k|x)$ of the true class k, using Stochastic Gradient Descent (SGD).

4.3 Regularization via Image Augmentation

Regularization techniques are essential in preventing overfitting and enhancing the generalization capabilities of deep models. Among these techniques, image augmentation stands out as a crucial method in supervised learning, well-known for its efficacy in regularization. While conventional augmentation methods like rotation, horizontal flips, and vertical flips are widely used, they often prove inadequate in domains characterized by limited data, such as medical imaging. Consequently, advanced approaches such as MixUp [50], CutMix [51], and ResizeMix [52] have received significant attention due to their ability to address the challenges posed by data scarcity and variability.These methods provide advanced solutions that extend beyond conventional techniques, facilitating the generation of diverse training samples and enhancing model robustness from various perspectives.

MixUp

MixUp generates synthetic samples by blending features and labels from two different images using weighted blending. Let (x_i, y_i) and (x_j, y_j) represent the features and labels of two images, respectively. MixUp combines these two images with weighted blending to create a new synthetic sample:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j$$

Here, $\lambda \in [0, 1]$ is a random weight value.

CutMix

CutMix creates synthetic samples by replacing regions of an image with patches from another image, controlled by a cut ratio (λ). Let (x_i, y_i) and (x_j, y_j) denote the features and labels of two images, respectively. CutMix blends these two images according to the cut ratio to create a new synthetic sample:

$$\tilde{x} = M \cdot x_i + (1 - M) \cdot x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j$$

Here, M is a mask matrix, and λ is a random cut ratio.

ResizeMix

ResizeMix merges images of different sizes by resizing them and then blending them using weighted blending. Let (x_i, y_i) and (x_j, y_j) represent the features and labels of two images, respectively. ResizeMix resizes one image to the size of the other and blends them with weighted blending to create a new synthetic sample:

$$\tilde{x} = \lambda x_i + (1 - \lambda)\text{resize}(x_j, \text{size}(x_i))$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

Here, $\text{resize}(x_j, \text{size}(x_i))$ resizes x_j to the size of x_i, and $\lambda \in [0, 1]$ is a random weight value.

5 Experiments and Discussion

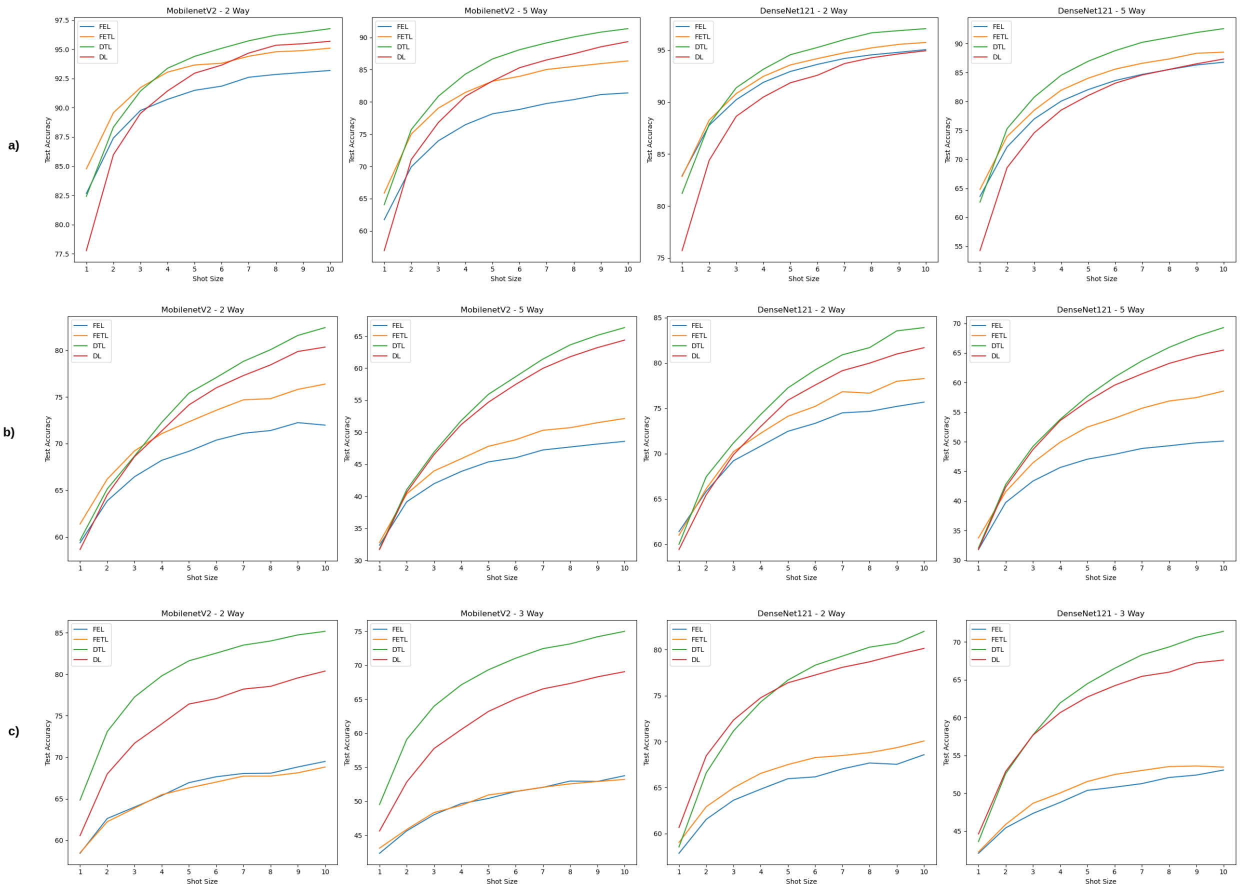

Figure 5: The graph compares the accuracies of the models based on the number of shots. It illustrates the results for a) SD-198, b) Derm7pt, and c) ISIC2018.

5.1 Implementation Details

The implementation of our study is carried out using the Python programming language with the PyTorch library. Nvidia 1080Ti GPU is employed during the model development. To ensure a fair comparison with the Meta-DermDiagnosis [8] model, we aligned the distinctions between base and novel classes in the datasets and ensured the similarity of deleted classes. All datasets are resized to 224x224x3 dimensions. ISIC2018 consists of 4 base and 3 novel classes, Derm7pt includes 13 base and 5 novel classes, and SD-198 encompasses 128 base and 70 novel classes. In contrast to the Meta-DermDiagnosis study, we partitioned the SD-198 base classes into training and validation sets. Classes with fewer

Table 1: Performance comparison of the proposed models on the **SD-198** skin lesion dataset for 2-way and 5-way classification tasks. Accuracy (%) values are presented.

Exp	**MobileNetV2**				**DenseNet121**			
	FEL	**FETL**	**DTL**	**DL**	**FEL**	**FETL**	**DTL**	**DL**
2W-1S	82.66	**84.77**	82.42	77.77	82.91	82.83	81.21	75.70
2W-2S	87.41	**89.57**	88.34	85.98	87.77	88.26	87.88	84.39
2W-3S	89.77	**91.73**	91.43	89.52	90.23	90.82	91.38	88.63
2W-4S	90.72	93.05	**93.38**	91.44	91.90	92.47	93.16	90.48
2W-5S	91.49	93.66	94.40	92.96	92.95	93.58	**94.56**	91.86
2W-6S	91.85	93.81	95.09	93.65	93.63	94.19	**95.26**	92.58
2W-7S	92.61	94.41	95.73	94.68	94.20	94.75	**96.02**	93.71
2W-8S	92.85	94.79	96.21	95.35	94.55	95.22	**96.67**	94.27
2W-9S	93.02	94.89	96.46	95.48	94.79	95.56	**96.88**	94.64
2W-10S	93.19	95.10	96.77	95.69	95.05	95.74	**97.07**	94.94
5W-1S	61.73	**65.85**	64.06	56.93	63.61	64.81	62.61	54.25
5W-2S	69.93	75.02	**75.72**	71.07	72.15	73.95	75.30	68.56
5W-3S	73.97	79.04	**80.88**	76.81	76.97	78.47	80.73	74.57
5W-4S	76.47	81.48	84.32	80.88	80.06	81.95	**84.53**	78.51
5W-5S	78.16	83.22	86.67	83.24	82.05	84.02	**86.95**	81.03
5W-6S	78.84	83.97	88.10	85.31	83.62	85.59	**88.78**	83.16
5W-7S	79.76	85.04	89.17	86.52	84.68	86.60	**90.22**	84.58
5W-8S	80.36	85.50	90.09	87.47	85.52	87.35	**91.05**	85.55
5W-9S	81.14	85.94	90.83	88.55	86.30	88.32	**91.91**	86.50
5W-10S	81.40	86.36	91.37	89.35	86.77	88.51	**92.57**	87.33

Table 2: Performance comparison of the proposed models on the **Derm7pt** skin lesion dataset for 2-way and 5-way classification tasks. Accuracy (%) values are presented.

Exp	**MobileNetV2**				**DenseNet121**			
	FEL	**FETL**	**DTL**	**DL**	**FEL**	**FETL**	**DTL**	**DL**
2W-1S	59.37	61.38	59.65	58.65	**61.40**	60.99	60.00	59.42
2W-2S	63.87	66.20	65.15	64.53	65.83	66.13	**67.45**	65.46
2W-3S	66.45	69.22	68.67	68.60	69.21	70.21	**71.17**	69.88
2W-4S	68.21	71.07	72.27	71.36	70.83	72.24	**74.31**	72.99
2W-5S	69.18	72.34	75.41	74.15	72.46	74.12	**77.25**	75.89
2W-6S	70.37	73.57	77.07	75.97	73.34	75.21	**79.23**	77.56
2W-7S	71.11	74.69	78.81	77.29	74.50	76.83	**80.90**	79.15
2W-8S	71.40	74.82	80.07	78.44	74.66	76.67	**81.69**	79.98
2W-9S	72.25	75.81	81.59	79.87	75.21	77.98	**83.54**	80.99
2W-10S	71.98	76.38	82.43	80.34	75.68	78.28	**83.90**	81.68
5W-1S	32.38	32.79	31.80	31.70	31.78	**33.74**	32.04	31.80
5W-2S	39.16	40.39	41.06	40.65	39.73	41.54	**42.78**	42.34
5W-3S	41.97	43.96	46.89	46.51	43.35	46.44	**49.19**	48.67
5W-4S	43.88	45.87	51.82	51.19	45.64	49.92	**53.77**	53.57
5W-5S	45.38	47.81	55.92	54.69	47.05	52.48	**57.68**	56.83
5W-6S	46.02	48.83	58.67	57.47	47.87	53.96	**60.92**	59.56
5W-7S	47.24	50.31	61.42	59.97	48.86	55.65	**63.65**	61.45
5W-8S	47.70	50.71	63.63	61.78	49.31	56.87	**65.93**	63.22
5W-9S	48.16	51.49	65.12	63.20	49.81	57.46	**67.80**	64.51
5W-10S	48.58	52.15	66.33	64.38	50.11	58.56	**69.28**	65.48

Table 3: Performance comparison of the proposed models on the **ISIC2018** skin lesion dataset for 2-way and 3-way classification tasks. Accuracy (%) values are presented.

Exp	**MobileNetV2**				**DenseNet121**			
	FEL	**FETL**	**DTL**	**DL**	**FEL**	**FETL**	**DTL**	**DL**
2W-1S	58.42	58.49	**64.83**	60.55	57.86	59.04	58.55	60.66
2W-2S	62.62	62.23	**73.10**	67.99	61.55	62.92	66.60	68.46
2W-3S	63.99	63.86	**77.26**	71.69	63.63	64.98	71.16	72.34
2W-4S	65.38	65.49	**79.81**	74.02	64.83	66.55	74.31	74.77
2W-5S	66.94	66.31	**81.63**	76.42	65.97	67.51	76.69	76.40
2W-6S	67.65	67.01	**82.55**	77.07	66.17	68.27	78.31	77.24
2W-7S	68.05	67.73	**83.52**	78.22	67.04	68.49	79.30	78.08
2W-8S	68.08	67.73	**84.02**	78.55	67.68	68.81	80.27	78.68
2W-9S	68.83	68.13	**84.75**	79.57	67.54	69.35	80.72	79.45
2W-10S	69.50	68.82	**85.18**	80.39	68.58	70.07	81.99	80.14
3W-1S	42.33	43.09	**49.51**	45.61	42.08	42.25	43.63	44.63
3W-2S	45.64	45.85	**59.10**	52.83	45.44	45.89	52.61	52.89
3W-3S	48.02	48.33	**63.98**	57.75	47.35	48.68	57.70	57.67
3W-4S	49.65	49.39	**67.13**	60.55	48.80	50.05	61.93	60.65
3W-5S	50.41	50.92	**69.35**	63.22	50.39	51.56	64.47	62.72
3W-6S	51.43	51.46	**71.06**	65.06	50.81	52.49	66.50	64.20
3W-7S	52.04	52.06	**72.47**	66.54	51.28	53.02	68.27	65.44
3W-8S	52.97	52.56	**73.16**	67.33	52.09	53.53	69.32	65.99
3W-9S	52.91	52.90	**74.21**	68.31	52.41	53.61	70.60	67.21
3W-10S	53.76	53.21	**75.01**	69.07	53.07	53.46	71.38	67.59

than 20 data points are designated as novel, while those with 20-30 data points are assigned to the validation set. To ensure a standardized testing environment for the four differently trained models, we employed a seed and deterministic mode, which worked in allowing each model to be tested on tasks of the same difficulty and order. Hyperparameters such as batch size used during testing were structured similarly for FETL and FEL models, with the only difference being the query set size set to 5 due to data insufficiency in novel classes. Data augmentation techniques are not applied to novel classes. All experiments are conducted using MobileNetV2 and DenseNet121 backboned models.

The sole distinction between the FEL (Few-Shot Episodic Learning) and FETL (Few-Shot Episodic Transfer Learning) models lies in the use of pretrained ImageNet weights for the backbone models in the FETL model. In contrast, the FEL model is trained on backbone models with random initialization. Data augmentation techniques such as RandomResizedCrop, RandomFlip, and ColorJitter are employed. In this framework, our models are trained with a 5-way 5-shot configuration, and all layers of the backbone models are fully opened for training.

Both DTL (Deep Transfer Learning) and DL (Deep Learning) models utilize ImageNet weights, but the DTL model is fine-tuned with the base classes of the datasets. The DL model, serving as a baseline, is chosen to compare performance without any fine-tuning, using ImageNet weights. In the DTL model, traditional training methods are employed instead of episodic training. For instance, in the case of the SD-198 dataset, 10% of the 128-class base section are allocated as the validation set. The MobileNetV2 model with ImageNet weights are then trained, aiming for the model to learn generalized features from the dataset. During testing, this model is used with ProtoNet in an episodic structure to evaluate and compare its performance.

Various augmentation techniques are employed in different training iterations of the DTL model to compare their effects and success rates. While the DTL-base model utilizes RandomResizeCrop and 50% Horizontal RandomFlip in the SD-198 and ISIC2018 datasets, Resize and 45% Horizontal and Vertical RandomFlip are used in the Derm7Pt dataset. Subsequently, the DTL-Base section is kept constant, and batch augmentation techniques such as CutMix, MixUp, and ResizeMix are added for comparison. Each added technique is labeled in the result tables as DTL-CutMix or DTL-ResizeMix. Our model named DTL-All-Augment represents a comprehensive model incorporating all three techniques on top of DTL-Base.

5.2 Experimental Analysis

In order to evaluate our framework and assess the effectiveness of our proposed methods, we performed experiments using 2-Way 1-Shot to 10-Shot and 5-Way 1-Shot to 10-Shot setups. These tests are conducted using three datasets: SD-198, Derm7pt, and ISIC2018. The comprehensive results of our model tests are detailed in Tables 1, 2, and 3. Moreover, we executed additional experiments based on the parameters specified in Tables 4, 5, and 6. These experiments are designed to compare our model's performance against benchmarks set in prior research, such as SCAN [19], MetaMed [18], and PFEMed [25], using the datasets referred to in Section [9–11]. In all these tables, the highest accuracy values are highlighted in bold, and the second highest results are underlined.

The SD-198 dataset contains 198 classes and consists only of clinical images. In contrast, the Derm7pt dataset has 18 classes with a mix of clinical and dermoscopic images. The ISIC2018 dataset, with its 7 classes, mainly features dermoscopic images. Although each of these three datasets shows a long-tail distribution, their unique features affect how models perform. Additionally, the number of parameters and the performances of the backbone models also play essential role in the overall model performance. For instance, MobileNetV2 has approximately 3.4 million parameters, while DenseNet121 boasts 8.1 million parameters. Furthermore, while episodic learning is employed in FEL and FETL models, transfer learning based training is utilized in DTL and DL models, a distinction that significantly impacts model performance. As can be observed in Figure 5, in all three datasets, FETL models consistently outperform FEL models, while DTL models demonstrate superior performance compared to DL models. Except for some rare cases, depending on the number of shots and dataset conditions, our experimental results indicate that the most successful approach in the proposed framework is the deep transfer learning based DTL model. As the number of shots increases, the performance of transfer learning-based DTL and DL models improves.

When comparing our models without utilizing augmentation techniques, datasets containing relatively large number of classes, such as SD-198 and Derm7pt, exhibit similar behaviors. Episodic learning-based models tend to be more successful when the number of shots is low (i.e. 1 or 2 shots), while DTL and DL models show a tendency to outperform as the number of shots increases. For instance, with MobilNetV2 model on the SD-198 dataset, for a 5-Way 1-Shot scenario, the FETL model achieved 65.85%, whereas the DTL model achieved 64.06%. For a 5-Way 10-Shot scenario, the FETL model achieved 86.36%, whereas the DTL model achieved 91.37% (Table 1). The behaviour of the models are similar in Derm7pt dataset as well (Table 2). The reason behind this phenomenon lies in the adaptability of episodic learning, which performs better in scenarios with fewer shots due to the large number of classes. Additionally, clinical images inherently encapsulate various differences, making models trained in an episodic manner more inclined to

Table 4: Comparison of our models and the SOTA methods. Values in the table are F1-scores of the corresponding models on the **SD-198** dataset.

Method	**Backbone**	**2 Way**		**5 Way**	
		1 Shot	**5 Shot**	**1 Shot**	**5 Shot**
PCN	Conv4	70.78±1.61	85.87±1.12	45.59±1.03	65.70±1.02
SCAN		78.00±1.51	91.01±0.90	55.60±1.07	75.65±0.87
SCAN	Conv6	77.64±1.50	88.28±1.03	54.07±1.24	74.73±0.92
NCA	WRN-28-10	71.27±1.50	84.23±1.19	45.91±1.08	62.83±1.01
Baseline		76.64±1.56	89.66±0.97	52.54±1.11	74.71±0.96
S2M2_R		77.15±1.59	90.97±0.89	55.49±1.13	78.17±0.84
NegMargin		77.98±1.45	90.65±0.92	56.04±1.14	77.75±0.87
PT+NCM		78.86±1.47	90.90±0.93	56.91±1.11	78.12±0.88
PEM$_b$E-NCM		78.70±1.49	90.94±0.95	57.42±1.11	78.78±0.90
EASY		79.44±1.51	91.43±0.96	57.77±1.12	79.53±0.89
SCAN		81.21±1.46	92.08±0.85	58.75±1.14	81.43±0.77
DTL-Base(Ours)	MobileNetV2	80.54±0.53	94.10±0.29	61.32±0.40	86.04±0.26
DTL-CutMix(Ours)		81.10±0.53	94.51±0.28	62.78±0.41	87.05±0.26
DTL-MixUp(Ours)		80.34±0.53	94.20±0.28	61.71±0.41	86.61±0.26
DTL-All-Augment(Ours)		81.44±0.53	94.76±0.28	63.22±0.41	87.48±0.25
DTL-ResizeMix(Ours)		**83.06±0.51**	**95.05±0.27**	**65.40±0.40**	**88.08±0.25**

Note:(1) The results of the SOTA models are taken from the SCAN [19]. (2) The complete references for the mentioned works in the table are as follows: PCN [14], SCAN [19], NCA [53], Baseline [47], S2M2_R [46], NegMargin [54], PT+NCM [48], PEM $_b$E-NCM [53], EASY [55].

tolerate these diversities. As the number of shots increases, models trained non-episodically become more successful in classifications.

Since the ISIC2018 dataset comprises dermoscopic data, the discriminative power of domain-specific features becomes more crucial. Hence, transfer learning-based DTL and DL models outperform FETL and FEL models. For instance, for MobileNetV2, the accuracy rates are 58.49% for 2-Way 1-Shot with FETL and 64.83% with DTL, and 68.82% for 2-Way 10-Shot with FETL and 85.18% with DTL (Table 3). Episodic learning-based methods tend to remain superficial in training. Additionally, in parallel with the SD-198 dataset, models tend to become more successful as the number of shots increases in ISIC2018 and Derm7p datasets. The MobileNetV2 and DenseNet121 models tend to yield complementary outcomes. DenseNet121, with its greater parameter count, is more susceptible to overfitting while training. Conversely, MobileNetV2, with its simpler architecture, is better suited for practical applications in this setting.

In all three datasets we studied, transfer learning methods generally perform better than episodic few-shot training in learning distinct features, except in cases where there is only one example provided. This observation suggests that even though episodic few-shot training is a more intricate approach, it may not be as effective as transfer learning in most scenarios. The exception is when extremely limited data is available, which is when episodic few-shot training can be valuable. This finding is consistent with recent research ([21]) that questions the practicality of complex few-shot training methods.

5.3 Comparison with Current State-of-the-Art

We also included a thorough analysis aimed at understanding how different skin disease datasets, each with unique features and conditions, influence the training process. We arranged the datasets in our research in a manner similar to SCAN [19] and PFEMed [25] studies for a fair comparison of the model performances. Specifically, we examined the performance benchmarks set by the SCAN model on the SD-198 and Derm7pt datasets, and by the PFEMed model on the ISIC2018 dataset.

In our research with the SD-198 dataset, we enhanced our Deep Transfer Learning (DTL) model by integrating various augmentation techniques, and compared these models with other state-of-the-art models in the field, particularly focusing on the SCAN model (Table 4). The SCAN model, which incorporates an unsupervised cluster branch, has been previously contrasted with the Meta-DermDiagnosis study. Our comparison also extended to various transfer learning-based Few-Shot Learning (FSL) methods; including the models like NCA [53], Baseline [47], S2M2_R [46], NegMargin [54], PT+NCM [48], PEM$_b$E-NCM [53], and EASY [55]. Notably, the SCAN model reported better results than these approaches in tests on the SD-198 dataset. The EASY model, which achieved results comparable to the SCAN model, used an ensemble technique and augmented images with a random resize crop method. Our approach, however, is different; we employ the MobileNetV2 model, leveraging ImageNet pre-trained weights instead of the

Table 5: Comparison of our models and the SOTA methods. Values in the table are Accuracies (%) of the corresponding models on the **ISIC2018** dataset.

Method	**Backbone**	**2 Way**			**3 Way**		
		3 Shot	**5 Shot**	**10 Shot**	**3 Shot**	**5 Shot**	**10 Shot**
Meta-DermDiagnosis	Conv6	64.50	73.50	79,70	-	-	-
MetaMed - Transf. Learn.	Conv4	66.88	73.88	81.37	54.83	59.33	69.75
MetaMed - Normal Aug.		72.75	75.62	81.37	54.83	59.33	69.75
MetaMed - CutOut		70.37	77.62	81.87	55.5	65.41	69.75
MetaMed - MixUp		75.37	78.25	84.25	58.5	61.25	71
MetaMed - CutMix		73.25	76.87	80.62	58.66	61.5	66.5
PT-MAP	WRN	68.15	70.87	74.19	53.17	55.61	59.57
Baseline+		64.77	70.27	74.67	53.2	54.16	57.87
NegMargin		71.33	72.67	75.17	60.69	57.58	63.04
Baseline		68.77	71.03	76.97	56.8	59.2	65.22
PFEMed		**81.69**	**83.87**	85.14	**66.94**	69.78	73.81
DTL-Base(Ours)	MobileNetV2	77.26	81.63	85.18	63.98	69.35	75.01
DTL-CutMix(Ours)		77.71	81.79	85.97	64.37	69.86	75.73
DTL-MixUp(Ours)		79.02	82.95	86.4	66.84	71.15	76.48
DTL-ResizeMix(Ours)		78,02	82.75	86.69	65.56	70.87	**76.97**
DTL-All-Augment(Ours)		78,96	83.21	**86.83**	66.12	**71.28**	76.94

Note: (1) The results of the SOTA models are taken from the PFEMed [25]. (2) The complete references for the mentioned works in the table are as follows: Meta-DermDiagnosis [8], MetaMed [18], PT-MAP [48], Baseline [47], NegMargin [54], PFEMed [25].

Table 6: Comparison of our models and the SOTA methods. Values in the table are Accuracies (%) of the corresponding models on the **Derm7pt** dataset.

Method	**Backbone**	**2 Way**	
		1 Shot	**5 Shot**
PCN	Conv4	59.98±1.28	70.62±13
SCAN		61.42±1.49	72.58±1.28
Meta-DermDiagnosis	Conv6	61.8	76.9
SCAN		62.80±1.34	76.65±1.21
NCA	WRN-28-10	56.32±1.29	67.18±1.15
Baseline		59.43±1.34	74.28±1.14
S2M2_R		61.37±1.33	79.83±1.34
NegMargin		58.00±1.44	70.12±1.30
PT+NCM		60.92±1.68	74.33±1.48
PEMb E_NCM		60.40±1.72	72.63±1.48
EASY		61.02±1.67	75.98±1.41
SCAN		66.75±1.35	**82.57±1.13**
PFEMed		**71.15**	80.27
DTL-Base(Ours)	DenseNet121	60.00±0.39	77.25±0.41
DTL-CutMix(Ours)		61.57±0.41	78.65±0.40
DTL-MixUp(Ours)		62.56±0.40	81.00±0.39
DTL-ResizeMix(Ours)		61.09±0.38	79.18±0.39
DTL-All-Augment(Ours)		60.88±0.40	78.30±0.40

Note: (1) The results of the SOTA models are taken from the SCAN [19]. (2) The complete references for the mentioned works in the table are as follows: PCN [14], Meta-DermDiagnosis [8], SCAN [19], NCA [53], Baseline [47], S2M2_R [46], NegMargin [54], PT+NCM [48], PEM $_b$E-NCM [53], EASY [55], PFEMed [25].

typically used WRN-28-10 model. Our augmentation strategy is more varied; we combined traditional techniques like Random Flip and Random Resize Crop with advanced methods such as CutMix, MixUp, and ResizeMix (as used by MetaMed model). This approach enabled our models to exceed the existing state-of-the-art results. Even though we used simple transfer learning based methodology, with the data augmentations our model could efficiently learn and transfer feature embeddings from base classes to novel classes.

Similarly, the PFEMed study compared its results on the ISIC2018 dataset with benchmarks set by studies like MetaMed [18], Meta-DermDiagnosis, PT-MAP [48], Baseline, and NegMargin. The PFEMed approach involved a comprehensive model with a dual-encoder structure combined with a Variational Autoencoder, distinguishing between general and specific features, and achieving state-of-the-art results. In contrast, the MetaMed study explored models employing augmentation techniques (such as CutOut, CutMix, and MixUp) during episodic training with Reptile model, and compared these to Meta-DermDiagnosis. In this dataset, observing the potential success of our DTL model trained with MobileNetV2, we applied similar augmentation techniques to the training data. The experiment results depict that our model performs comparable results with the state-of-the-art models; producing second best results in 3-Shot and 5-Shot cases. Moreover, we obtain best results when the number of samples is increased (i.e. 10-Shot case) (Table 5).

We also compared our models with leading models on the Derm7pt dataset. The unique features of this dataset played a key role in improving performance, especially the sub-cluster aspect in the SCAN model. We included findings from the PFEMed study in our comparison. By using the DTL-MixUp model, based on the DenseNet121 framework, we achieved results that align with existing benchmarks in the 2-Way 5-Shot scenario, where our model ranked second in accuracy. Similar to what we observed in our early experiments without data augmentations, in one-shot settings, the DTL model didn't perform as well compared to other models (see Figure 5). It's important to note that our method relies solely on transfer learning techniques and enhances the pretrained ImageNet model with data augmentations. These outcomes indicate that this fundamental approach remains effective in managing long-tail distributions, without the need for intricate training methods.

6 Conclusion and Future Work

In this study, we explore the effectiveness of episodic and traditional training methods combined with few-shot learning through transfer learning for classifying rare skin diseases in situations with limited data. Current literature emphasize that datasets with long-tail distributions present specific challenges, highlighting the necessity for specialized training approaches to achieve successful outcomes. But there is a lack of research combining basic few-shot training models with well-known transfer learning methodologies in this domain. Our study seeks to fill this gap by evaluating these established training methods on major datasets often used as benchmarks for long-tail distributions in skin diseases. In this context, we presented evaluations of three publicly available skin image datasets using the FETL, FEL, DTL, and DL models. Notably, the DTL model, which integrates transfer learning with conventional training techniques, consistently outperformed others in these datasets. By enhancing our models with MixUp, CutMix, and ResizeMix techniques, as recommended in existing literature, we were able to reach leading performance on the SD-198 and comparable performances with the state-of-the-art methods in ISIC2018 and Derm7pt datasets.

These findings underline the effectiveness of transfer learning, a well-established method, in the context of few-shot learning for long-tail distributions in skin disease classification. We suggest that further improvements in accuracy and robustness can be achieved through expanded data augmentation and exploring different model architectures. Based on the insights gained from this study, we believe that further research is crucial, especially for dermoscopic datasets facing data scarcity. Moreover, applying these findings to a broader range of medical imaging datasets, including X-rays, CT scans, and MRIs, could be helpful in detecting and diagnosing new diseases.

References

[1] Duowen Chen, Yunhao Bai, Wei Shen, Qingli Li, Lequan Yu, and Yan Wang. Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23869–23878, 2023.

[2] Bo Wang, Qian Li, and Zheng You. Self-supervised learning based transformer and convolution hybrid network for one-shot organ segmentation. *Neurocomputing*, 527:1–12, 2023.

[3] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7422–7432, 2023.

[4] Indrajit Mazumdar and Jayanta Mukherjee. Fully automatic mri brain tumor segmentation using efficient spatial attention convolutional networks with composite loss. *Neurocomputing*, 500:243–254, 2022.

[5] Suraj Mishra, Yizhe Zhang, Li Zhang, Tianyu Zhang, X Sharon Hu, and Danny Z Chen. Data-driven deep supervision for skin lesion classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pages 721–731. Springer, 2022.

[6] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

[7] Brian Hon Yin Chung, Jeffrey Fong Ting Chau, and Gane Ka-Shu Wong. Rare versus common diseases: a false dichotomy in precision medicine. *NPJ Genomic Medicine*, 6(1):19, 2021.

[8] Kushagra Mahajan, Monika Sharma, and Lovekesh Vig. Meta-dermdiagnosis: Few-shot skin disease identification using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 730–731, 2020.

[9] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 206–222. Springer, 2016.

[10] 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.

[11] 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.

[12] Md Kamrul Hasan, Md Toufick E Elahi, Md Ashraful Alam, Md Tasnim Jawad, and Robert Martí. Dermoexpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation. *Informatics in Medicine Unlocked*, 28:100819, 2022.

[13] Satin Jain, Udit Singhania, Balakrushna Tripathy, Emad Abouel Nasr, Mohamed K Aboudaif, and Ali K Kamrani. Deep learning-based transfer learning for classification of skin cancer. *Sensors*, 21(23):8142, 2021.

[14] Viraj Prabhu, Anitha Kannan, Murali Ravuri, Manish Chaplain, David Sontag, and Xavier Amatriain. Few-shot learning for dermatological disease diagnosis. In *Machine Learning for Healthcare Conference*, pages 532–552. PMLR, 2019.

[15] Xiaomeng Li, Lequan Yu, Yueming Jin, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Difficulty-aware meta-learning for rare disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 357–366. Springer, 2020.

[16] Delong Zhang, Mengqun Jin, and Peng Cao. St-metadiagnosis: Meta learning with spatial transform for rare skin disease diagnosis. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2153–2160. IEEE, 2020.

[17] Wei Zhu, Haofu Liao, Wenbin Li, Weijian Li, and Jiebo Luo. Alleviating the incompatibility between cross entropy loss and episode training for few-shot skin disease classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 330–339. Springer, 2020.

[18] Rishav Singh, Vandana Bharti, Vishal Purohit, Abhinav Kumar, Amit Kumar Singh, and Sanjay Kumar Singh. Metamed: Few-shot medical image classification using gradient-based meta-learning. *Pattern Recognition*, 120:108111, 2021.

[19] Shuhan Li, Xiaomeng Li, Xiaowei Xu, and Kwang-Ting Cheng. Dynamic subcluster-aware network for few-shot skin disease classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[20] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 124–141. Springer, 2020.

[21] Steinar Laenen and Luca Bertinetto. On episodes, prototypical networks, and few-shot learning. *Advances in Neural Information Processing Systems*, 34:24581–24592, 2021.

[22] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020.

[23] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 403–412, 2019.

[24] Junsheng Xiao, Huahu Xu, DiKai Fang, Chen Cheng, and HongHao Gao. Boosting and rectifying few-shot learning prototype network for skin lesion classification based on the internet of medical things. *Wireless Networks*, 29(4):1507–1521, 2023.

[25] Zhiyong Dai, Jianjun Yi, Lei Yan, Qingwen Xu, Liang Hu, Qi Zhang, Jiahui Li, and Guoqiang Wang. Pfemed: Few-shot medical image classification using prior guided feature enhancement. *Pattern Recognition*, 134:109108, 2023.

[26] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[28] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.

[29] Wei Ying, Yu Zhang, Junzhou Huang, and Qiang Yang. Transfer learning via learning to transfer. In *International Conference on Machine Learning*, pages 5085–5094. PMLR, 2018.

[30] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

[31] Amirreza Mahbod, Gerald Schaefer, Chunliang Wang, Georg Dorffner, Rupert Ecker, and Isabella Ellinger. Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Computer methods and programs in biomedicine*, 193:105475, 2020.

[32] Zhiwei Qin, Zhao Liu, Ping Zhu, and Yongbo Xue. A gan-based image synthesis method for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 195:105568, 2020.

[33] Lina Liu, Lichao Mou, Xiao Xiang Zhu, and Mrinal Mandal. Automatic skin lesion classification based on mid-level feature learning. *Computerized Medical Imaging and Graphics*, 84:101765, 2020.

[34] Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Alain Pitiot, and Chunliang Wang. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics*, 71:19–29, 2019.

[35] Sema Atasever, Nuh Azginoglu, Duygu Sinanc Terzi, and Ramazan Terzi. A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. *Clinical Imaging*, 94:18–41, 2023.

[36] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[37] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.

[38] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[39] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[40] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[41] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[42] Xiaoxu Li, Xiaochen Yang, Zhanyu Ma, and Jing-Hao Xue. Deep metric learning for few-shot image classification: A review of recent developments. *Pattern Recognition*, page 109381, 2023.

[43] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE international conference on computer vision*, pages 3018–3027, 2017.

[44] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[45] Ramnath Kumar, Tristan Deleu, and Yoshua Bengio. The effect of diversity in meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8396–8404, 2023.

[46] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2218–2227, 2020.

[47] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[48] Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Leveraging the feature distribution in transfer-based few-shot learning. In *International Conference on Artificial Neural Networks*, pages 487–499. Springer, 2021.

[49] Wei Zhu, Wenbin Li, Haofu Liao, and Jiebo Luo. Temperature network for few-shot learning with distribution-aware large-margin metric. *Pattern Recognition*, 112:107797, 2021.

[50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[51] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[52] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101*, 2020.

[53] Zhirong Wu, Alexei A Efros, and Stella X Yu. Improving generalization via scalable neighborhood component analysis. In *Proceedings of the european conference on computer vision (ECCV)*, pages 685–701, 2018.

[54] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 438–455. Springer, 2020.

[55] Yassir Bendou, Yuqing Hu, Raphael Lafargue, Giulia Lioi, Bastien Pasdeloup, Stéphane Pateux, and Vincent Gripon. Easy—ensemble augmented-shot-y-shaped learning: State-of-the-art few-shot classification with simple components. *Journal of Imaging*, 8(7):179, 2022.