Meta-Transfer Derm-Diagnosis: Exploring Few-Shot Learning and Transfer Learning for Skin Disease Classification in Long-Tail Distribution

Zeynep ÖZDEMİR*, Hacer YALIM KELEŞ, Ömer Özgür TANRIÖVER

***Abstract*— Building accurate models for rare skin diseases remains challenging due to the lack of sufficient labeled data and the inherently long-tailed distribution of available samples. These issues are further complicated by inconsistencies in how datasets are collected and their varying objectives. To address these challenges, we compare three learning strategies: episodic learning, supervised transfer learning, and contrastive self-supervised pretraining, within a few-shot learning framework. We evaluate five training setups on three benchmark datasets: ISIC2018, Derm7pt, and SD-198. Our findings show that traditional transfer learning approaches, particularly those based on MobileNetV2 and Vision Transformer (ViT) architectures, consistently outperform episodic and self-supervised methods as the number of training examples increases. When combined with batch-level data augmentation techniques such as MixUp, CutMix, and ResizeMix, these models achieve state-of-the-art performance on the SD-198 and Derm7pt datasets, and deliver highly competitive results on ISIC2018. All the source codes related to this work will be made publicly available soon at the provided URL.**

***Index Terms*— Few-shot learning, Long-tail distribution, Medical image classification, Skin disease classification, Self Supervised learning, Transfer learning, Explainability, Uncertainty estimation**

I. 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, particularly certain skin conditions, presents unique challenges specific to the domain. Skin disease datasets, unlike those for natural image classification, often exhibit long-tailed distributions, where a few common classes dominate while rare classes are underrepresented. Additionally, variations in image quality and the presence of artifacts such as hair, rulers, and ink markings further complicate the classification process. The diversity of imaging modalities, including clinical and dermoscopic images, adds another layer of complexity, as these modalities vary in resolution, feature representation, and clinical relevance. Anatomical and biological diversity, subtle visual distinctions, and patient demographics introduce further complexities that directly impact model performance. Moreover, the need for expert clinicians to annotate datasets and strict data privacy regulations that limit data sharing represent significant obstacles in this field. These challenges underscore the necessity for developing advanced, domain-specific methodologies designed to address the complexities of skin disease classification [8]–[11].

Given these limitations, Few-shot learning (FSL) has emerged as a promising approach to address the challenges of class imbalance and data scarcity in medical imaging. FSL offers versatile solutions in image classification, detection,

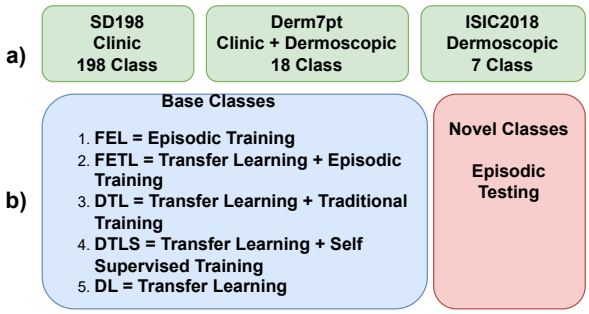

Fig. 1. Skin Lesion Classification Framework: a) Overview of benchmark datasets, emphasizing long-tailed distribution, class counts, and modality differences. b) Framework for analyzing transfer learning's impact on training strategies.

Manuscript received April 24, 2024. (Corresponding author: Zeynep ÖZDEMİR)

Zeynep ÖZDEMİR is with the Department of Computer Engineering, Graduate School of Natural and Applied Sciences, Ankara University, Ankara, Turkey. (e-mail: zynpozdemir@ankara.edu.tr).

Hacer YALIM KELEŞ is with the Department of Computer Engineering, Hacettepe University, Ankara, Turkey. (e-mail: hacerkeles@hacettepe.edu.tr).

Ömer Özgür TANRIÖVER is with the Department of Computer Engineering, Ankara University, Ankara, Turkey. (e-mail: tanriover@ankara.edu.tr).

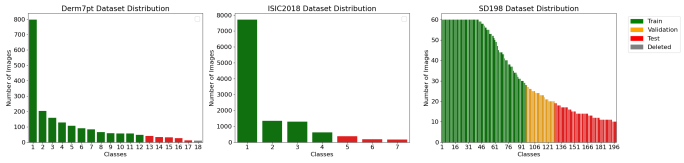

Fig. 2. The figure shows the distributions of the datasets SD-198 [12], Derm7pt [13], and ISIC2018 [14], highlighting their long-tailed nature with some classes having very few instances. Base classes (common diseases) are marked as train (green) and validation (yellow), while novel classes (rare diseases) are labeled as test (red). In the Derm7pt dataset, classes with very few examples are shown as deleted (grey) and excluded from use.

and segmentation tasks across both natural and medical images [15]–[23]. Specifically, FSL methods have demonstrated notable effectiveness in navigating the complexities of skin disease classification [11].

Building upon these advancements, 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 [24], [25], those relying on few-shot learning [11], [26]–[31], and approaches using cross-domain few-shot learning [32]. The state of the art models in this domain, such as Meta-DermDiagnosis, MetaMed, and PCN models [11], [26], [30] are designed to extract and learn high-level, domain-specific features during their training process.

In their study on the ISIC2018 dataset, Li et al. (2020) introduced the Difficulty Aware Meta-Learning (DAML) model, addressing task variability challenges [27]. Similarly, Mahajan et al. (2020) developed the Meta-DermDiagnosis model, experimenting with the SD-198, Derm7pt, and ISIC2018 datasets [11]. This model incorporated Prototypical Networks and Reptile to enhance robustness but faced criticism for its reliance on symmetric dataset orientations. 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 [30]. Both MetaMed and Meta-DermDiagnosis models aim to enhance feature representation by broadening data augmentation and diversification.

The mentioned studies in the field of skin disease classification have employed episodic learning as a method to acquire knowledge transferable to new classes. This approach involves dividing learning problems into small training and validation subsets to simulate scenarios encountered during evaluation. However, some studies in the FSL domain working with natural images have criticized episodic learning, arguing that its constraints are unnecessary and that using training groups in this manner is data-inefficient [33]. Similarly, [34] argued that, in tasks involving rare classes, the effectiveness of rapid adaptation depends more on the quality of the learned representation than on the few-shot learning algorithm itself. To enhance representation quality, they incorporated self-supervised learning components into their models. In a similar context, [35] proposed the concept of Meta-Transfer Learning, which aims to refine the learned representation. These studies conducted various experiments comparing the effectiveness of fine-tuning pre-trained Deep Neural Network (DNN) models with and without episodic learning. Their findings highlight 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 [11], [36], [37]. Although current methods show promise in specific dataset contexts, their ability to generalize broadly remains limited. Most research in this field has focused on developing dataset-specific solutions, addressing the unique challenges and sensitivities arising from variations in class counts, data formats, and other critical characteristics. On the other hand, episodic training techniques, which have been criticized for their inefficiency in the natural image domain of FSL, continue to be utilized in this context.

In this study, we aim to explore the effects of episodic learning, transfer learning and self-supervised learning for rare skin disease classification. One of our objective was to evaluate different model training approaches, using a benchmark test set specifically for rare skin diseases. In addition to transfer learning approaches using supervised ImageNet-pretrained models such as ResNet50 [38], DenseNet121 [39], MobileNetV2 [40], Vision Transformer (ViT) [41], ConvNeXt [42], and EfficientNet [43], we extended our evaluation to include models pretrained via contrastive self-supervised learning, including SimCLR [44], MoCo-v3 [45], DINO-v2 [46], SimSiam [47], SwAV [48], and ViT-MMF [49]. Additionally, we investigated the potential of transformer-based foundation models developed for medical imaging, such as BioViL [50] and MedCLIP [51].

In this context, we implemented five distinct training approaches, each evaluated using consistent metrics for novel classes (Figure 1). Our first approach, Few-Shot Episodic Learning (FEL), applies episodic training to models initialized with random weights (i.e., without any pre-training). The second, Few-Shot Episodic Transfer Learning (FETL), extends this by incorporating ImageNet-pretrained weights, which are fine-tuned on the long-tail skin disease dataset within an episodic learning setup. The third method, Deep Transfer Learning (DTL), omits episodic training and instead fine-tunes ImageNet-pretrained models directly on base classes. The fourth, Deep Transfer Learning with Self-Supervised Learning (DTLS), uses contrastively pretrained models, which are then fine-tuned with supervised training on the target dataset. Finally, Deep Learning (DL) serves as a baseline, where ImageNet-pretrained models are evaluated without any additional training or fine-tuning.

FETL, DTL, and DTLS were designed to incorprate FSL natural image domain advancements to address rare challenge. Additionally, performance improvements were targeted by integrating data augmentation techniques, such as MixUp, CutMix, and ResizeMix, specifically adapted for the unique properties of skin disease datasets.

To evaluate the effectiveness of explored methods we carried out extensive testing across three benchmark skin disease datasets: SD-198, Derm7pt, and ISIC2018. Our comprehensive analysis provide insights for effective strategies in tackling the

challenge of long-tailed data distribution in skin diseases. The key contributions of our study are summarized as follows:

- We evaluate different model training approaches, using a benchmark test set specifically curated for long-tail distributions in rare skin diseases. The final comparison 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.rare skin disease.
- 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 systematically evaluated a wide range of pretrained and foundation models (e.g., ConvNeXt, ViT, ResNet50, MedCLIP, BioVIL) under few-shot settings, both with and without additional fine-tuning. Our results provide valuable insights into selecting strong initialization points for rare skin disease classification, emphasizing the importance of pretraining choices in low-data regimes.
- We demonstrated that combining transfer learning and self-supervised learning with few-shot learning significantly enhances both the learned representation and the testing performance in the context of rare skin diseases.

II. Related Works

A. 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 pretrained on extensive datasets have demonstrated superior generalization performance compared to models initialized randomly [52]. Various techniques are employed to facilitate the transfer of knowledge between different source-target domains [53]–[56].

For skin disease classification, transfer learning has been widely utilized. Architectures such as EfficientNet, ResNet, and DenseNet, typically pre-trained on ImageNet or dermatological datasets, have been transferred to datasets like ISIC2017 and ISIC2018 [57]–[60], demonstrating notable performance gains in data-scarce scenarios. A more comprehensive literature discussion is provided in [61].

Beyond conventional supervised pretraining on foundation models, recent advances in self-supervised learning (SSL) have introduced noval mechanisms for learning transferable and robust representations without relying on labels. Previous studies in the few-shot skin disease classification domain, such as MetaMed [30], have explored episodic training combined with data augmentation, using shallow architectures and transfer learning on the ISIC2018 dataset. In contrast, our approach leverages deeper architectures to enhance feature representation capacity, aiming to better capture the complexity of rare skin disease patterns.

B. 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 [62]–[64]. 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 [65] utilizes Euclidean distance, Matching Networks [66] employs cosine similarity, and Relation Networks [67] utilizes its own CNN-based measurement module for this purpose [68]. 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 [69], [70].

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 [33], [34]. Additionally, [71] 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 [72], Baseline [73], PT-MAP [74], and Meta-Transfer Learning [35] 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.

C. 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, [27] 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 [30], 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. [29] advocated for the superiority of the Query-Relative loss over the Cross-Entropy loss commonly used in FSL. Additionally, [75] 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. [11] 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, the study in [37] 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, [36] 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, [31] 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.

Building upon previous studies, we present a comprehensive framework for few-shot rare skin disease classification by systematically evaluating a wide range of backbone architectures and training strategies. Specifically, we experimented with architectures including ResNet50 [38], DenseNet121 [39], MobileNetV2 [40], ViT [41], ConvNeXt [42], and EfficientNet [43], most of which were pretrained using supervised learning on ImageNet. To explore alternative pretraining paradigms, we also employed contrastive self-supervised methods such as SimCLR [44], MoCo-v3 [45], SwAV [48], and SimSiam [47], as well as distillation-based approaches like DINOv2 [46] and ViT-MMF [49], particularly on ResNet50 and ViT backbones. Additionally, transformer-based foundation models tailored for medical imaging, including BioViL [50] and MedCLIP [51], were included as part of our extended transfer learning evaluation.

Our comparative analysis focuses on three major training paradigms: traditional supervised transfer learning, contrastive self-supervised pretraining, and episodic few-shot learning. While contrastive pretraining showed benefits in certain cases, our results indicate that supervised transfer learning generally yields more robust performance in few-shot settings. Furthermore, episodic learning methods often failed to fully exploit the available data. To improve generalization, we incorporated augmentation techniques such as MixUp, CutMix, and ResizeMix, customized to address the characteristics of dermatological datasets. Importantly, our evaluations rely solely on existing benchmark datasets and do not require any supplementary external data, making the findings potentially applicable to other low-resource rare disease classification tasks. Detailed results and comparative insights are presented in Section V-D.

III. Datasets and Evaluation

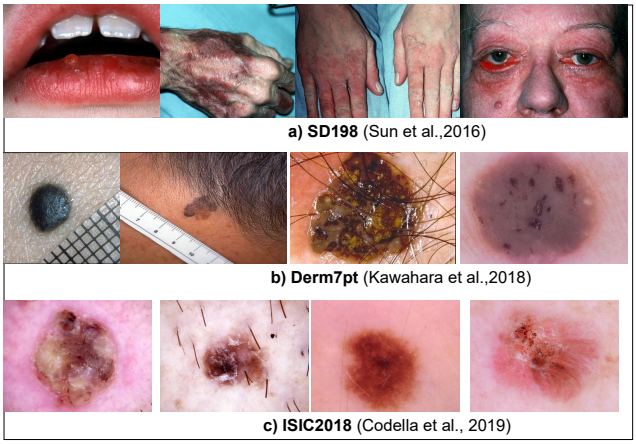

Fig. 3. Some sample images from skin disease classification datasets

The SD-198 dataset [12] 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 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 [11] and SCAN [31], 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 2, and sample images are provided in Figure 3-a.

The Derm7pt dataset [13] 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 3-b. The dataset distribution is also shared in Figure 2.

The ISIC 2018 Skin Lesion dataset [14] 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 [11] and PFEMed [37] 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 3-c for exemplar images drawn from the dataset. For data distribution, please see Figure 2.

IV. 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 five distinct model training strategies that are particularly valuable in scenarios with limited data samples. These strategies include Few-Shot Episodic Transfer Learning (FETL), Few-Shot Episodic Learning (FEL), Deep Transfer Learning (DTL), Deep Transfer Learning with Self-Supervised Adaptation (DTLS), and Standard Deep Learning (DL).

We implement a precise evaluation process, employing a combination of the selected benchmark datasets (Section III), 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.

A. 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 [68] 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. The classifier should then be able to accurately categorize the 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 4, we maintained a consistent evaluation by keeping D_{novel} fixed. Here, D_{train} denotes the datasets used during training. We formulated five 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 and DTLS, 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 summarized in Figure 5, our FETL, FEL, DTLS 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 five proposed analyses include adapted few-shot training and testing methodologies considering the data extracted carefully from the same long-tail distributions (FETL, FEL, DTL and DTLS), while the last one (DL) corresponds to a cross-domain evaluation.

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

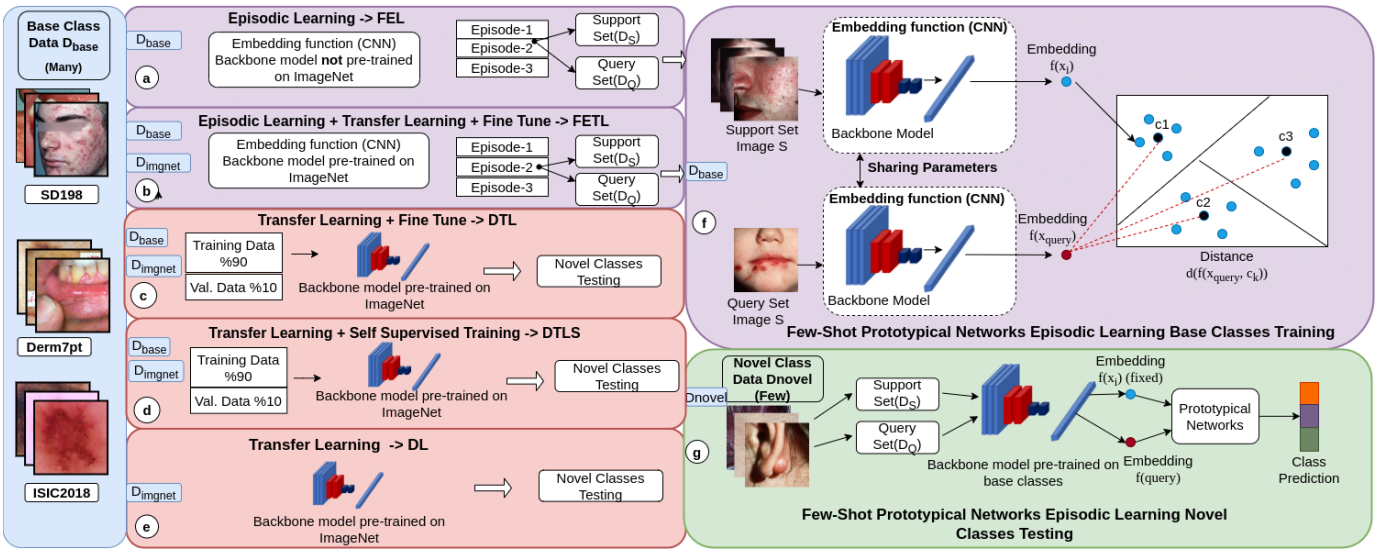

Fig. 4. Overall framework of our pipeline: Meta-Transfer Derm-Diagnosis. a) Episodic learning is combined with DenseNet and MobileNet architectures without the use of ImageNet weights. b) ImageNet pre-trained weights are utilized along with the application of an episodic learning strategy. c) Pre-trained weights and all base class data are employed for fine-tuning with ImageNet. d) Pre-trained weights and all base class data are employed for self-supervised training with ImageNet. e) 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. g) Common evaluation scheme using novel data across segments a, b, c, d and e.

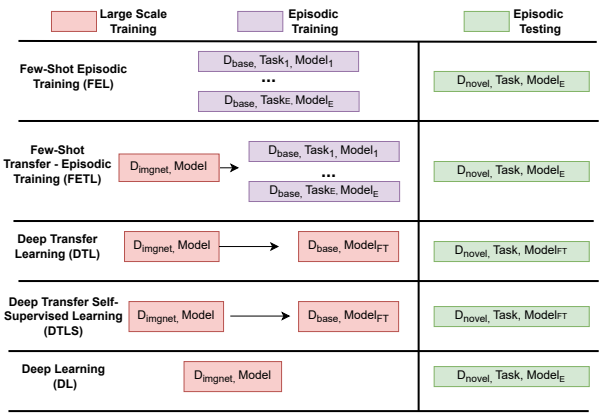

Fig. 5. The flowchart illustrating the components and the training strategies of the FEL, FETL, DTL, DTLS and DL models within the proposed framework.

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

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.

B. Prototypical Networks

As depicted in Figure 4 (top right section), we used Prototypical Networks [65] 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:

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

The network utilizes a distance function $d : \mathbb{R}^M \times \mathbb{R}^M \rightarrow [0, +\infty)$ 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:

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

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

C. 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 [76], CutMix [77], and ResizeMix [78] 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 perspective.

V. Experiments and Discussion

A. 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 [11] 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 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. A diverse set of backbone architectures, including convolutional and transformer-based models, was used throughout the experiments, reflecting both conventional and self-supervised transfer learning settings.

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.

Similarly, both DTL (Deep Transfer Learning), DTLS (Deep Transfer Learning with Self-Supervised Adaptation) and DL (Deep Learning) models utilize ImageNet weights, but the DTL and DTLS 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 and DTLS setups, a standard training approach is utilized instead of episodic training; for instance, in SD-198, 10% of the base classes are set aside for validation during fine-tuning. DTLS models were pretrained in a self-supervised manner using contrastive loss, and subsequently evaluated by attaching a linear classification head and fine-tuning the models on base classes. The results of this evaluation are presented in Table II. For models based on Vision Transformer (ViT-Base), fine-tuning was performed using Low-Rank Adaptation (LoRA) to enable efficient parameter updates. For evaluation consistency, all models,including DTL, DTLS, and DL,were tested using a prototypical episodic structure, mimicking few-shot scenarios during inference.

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.

B. Experimental Analysis

In order to evaluate our framework and assess the effectiveness of the methods, we conducted experiments using five few-shot configurations: 2-Way 1-Shot (2W1S), 2-Way 5-Shot (2W5S), 2-Way 10-Shot (2W10S), 5-Way 1-Shot (5W1S), and 5-Way 5-Shot (5W5S). These tests are conducted using three datasets: SD-198, Derm7pt, and ISIC2018. Model names in the results Table-II follow a consistent notation where, for example, *ResNet50(SimCLR)* refers to a ResNet50 backbone that was first pretrained using the SimCLR self-supervised method and then fine-tuned on base classes for classification.

TABLE I
FEW-SHOT CLASSIFICATION ACCURACIES OF BACKBONE MODELS WITHOUT FINE-TUNING ON **SD-198**, **DERM7PT**, AND **ISIC2018** DATASETS.

Methods	Params (M)	Flops (G)	SD-198			Derm7pt			ISIC2018		
			2W1S	2W5S	5W5S	2W1S	2W5S	5W5S	2W1S	2W5S	3W5S
MedCLIP(ViT) [51]	27.94	4.5	59.28	64.42	36.6	57.88	67.71	31.16	50.74	51.76	35.09
MedCLIP(ResNet50) [51]	25.56	4.11	66.58	80.56	60.06	57.66	64.62	38.25	51.96	55.64	39.5
BioVIL(ResNet-ViT) [50]	25.13	4.2	61.51	72.25	47.96	56.46	62.52	36.8	50.95	52.82	36.49
ConvNeXt(Base) [42]	88.59	15.36	80.32	92.86	81.15	60.25	76.84	55.95	59.55	72.27	58.57
EfficientNetB5 [43]	30.39	10.8	72.88	85.56	70.12	56.0	64.45	40.05	62.14	69.91	54.35
ResNet50 [38]	25.56	4.12	76.77	92.39	82.15	58.45	74.21	55.06	59.97	76.49	63.61
EfficientNetB4 [43]	19.34	4.66	74.59	88.44	74.33	56.55	68.78	44.58	**62.91**	72.56	57.0
DenseNet121 [39]	7.98	2.88	75.7	91.86	81.03	59.42	75.89	56.83	60.66	76.4	62.72
MobileNetV2 [40]	3.5	0.32	77.77	**92.96**	**83.24**	58.65	74.15	54.69	60.55	76.42	63.22
MoCo-v3(ResNet50) [45]	68.01	4.11	69.06	89.52	78.08	58.82	76.53	58.43	58.6	76.58	64.52
SimSiam(ResNet50) [47]	38.2	4.11	71.28	90.35	79.11	58.81	76.14	58.57	60.55	**78.05**	**65.74**
SwAV(ResNet50) [48]	28.35	4.11	70.24	89.57	77.8	57.69	74.95	56.64	58.51	76.88	64.88
SimCLR(ResNet50) [44]	27.97	4.11	71.55	89.9	77.93	58.26	74.47	56.34	59.14	75.11	62.14
MoCo-v3(ViT) [45]	215.68	17.58	75.58	85.77	70.08	**88.4**	92.18	65.24	53.13	55.98	40.18
DINOv2(ViT) [46]	86.58	152.0	70.44	79.03	57.63	86.89	90.07	57.97	54.02	58.04	42.21
ViT-MMF [49]	86.57	17.58	71.69	87.65	73.74	81.92	91.05	75.62	53.37	58.93	43.46
ViT(Base) [41]	86.57	17.58	**81.34**	92.55	82.14	86.85	**93.05**	**77.9**	52.87	58.28	42.31

TABLE II
FEW-SHOT CLASSIFICATION ACCURACIES OF FINE-TUNED MODELS ON **SD-198**, **DERM7PT**, AND **ISIC2018** DATASETS.

Methods	SD-198					Derm7pt					ISIC2018				
	2W1S	2W5S	2W10S	5W1S	5W5S	2W1S	2W5S	2W10S	5W1S	5W5S	2W1S	2W5S	2W10S	3W1S	3W5S
ResNet50+Aug DTL	78.53	94.47	97.52	59.52	87.24	61.62	78.86	86.71	34.03	60.41	61.7	78.28	83.47	46.72	65.86
ResNet50(SimCLR)+Aug DTLS	82.56	93.00	94.97	62.01	79.20	63.91	80.58	85.62	36.52	63.26	62.9	78.54	82.88	47.58	66.1
ResNet50(SimSiam)+Aug DTLS	71.28	90.35	95.08	47.09	76.01	58.81	76.14	83.13	31.19	58.57	60.65	78.05	82.40	45.05	65.74
ResNet50(MoCo-v3)+Aug DTLS	69.09	89.52	94.88	43.92	76.00	58.82	76.53	83.36	31.15	58.43	58.6	76.58	81.04	43.32	64.52
DenseNet121 FEL	82.91	92.95	95.69	63.61	82.05	61.4	72.46	75.68	31.78	47.05	57.86	65.97	68.58	42.08	50.39
DenseNet121 FETL	82.83	93.58	95.05	64.81	84.02	60.99	74.12	78.28	33.74	52.48	59.04	67.51	70.07	42.25	51.56
DenseNet121 DTL	81.21	94.56	95.74	62.61	86.95	60.0	77.25	83.9	32.04	57.68	58.55	76.69	81.99	43.63	64.47
DenseNet121+Aug DTL	82.29	**95.46**	**97.97**	65.54	**88.76**	62.56	81.00	87.16	34.75	61.91	56.67	65.06	67.85	39.81	48.66
MobileNetV2 FEL	82.66	91.49	93.19	61.73	78.16	59.37	69.18	71.98	32.38	45.38	58.42	66.94	69.5	42.33	50.41
MobileNetV2 FETL	**84.77**	93.66	95.10	65.85	83.22	61.38	72.34	76.38	32.79	47.81	58.49	66.31	68.82	43.09	50.92
MobileNetV2 DTL	82.42	94.4	96.77	64.06	86.67	59.65	75.41	82.43	31.8	55.92	64.83	81.63	85.18	49.51	69.35
MobileNetV2+Aug DTL	84.63	**95.46**	97.34	**67.94**	88.49	60.37	77.41	84.93	33.54	59.45	**65.55**	**83.21**	**86.83**	**50.40**	**71.28**
ViT(Base)+LoRA+Aug DTL	82.04	92.80	95.80	64.76	85.60	**86.35**	**93.68**	**95.56**	**66.33**	**80.74**	61.59	65.64	67.98	45.61	49.55

TABLE III
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 [26]	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 [31]	Conv6	77.64±1.50	88.28±1.03	54.07±1.24	74.73±0.92
NCA [79]	WRN-28-10	71.27±1.50	84.23±1.19	45.91±1.08	62.83±1.01
Baseline [73]		76.64±1.56	89.66±0.97	52.54±1.11	74.71±0.96
S2M2_R [72]		77.15±1.59	90.97±0.89	55.49±1.13	78.17±0.84
NegMargin [80]		77.98±1.45	90.65±0.92	56.04±1.14	77.75±0.87
PT+NCM [74]		78.86±1.47	90.90±0.93	56.91±1.11	78.12±0.88
PEM_bE-NCM [79]		78.70±1.49	90.94±0.95	57.42±1.11	78.78±0.90
EASY [81]		79.44±1.51	91.43±0.96	57.77±1.12	79.53±0.89
SCAN [31]		81.21±1.46	92.08±0.85	58.75±1.14	81.43±0.77
DTL+Aug(Ours)	ResNet50	75.04±0.38	94.11±0.62	55.93±0.43	86.55±0.71
DTL+Aug(Ours)	ViT(Base)	81.76±0.65	92.10±0.45	64.00±0.37	85.10±0.59
DTL+Aug(Ours)	DenseNet121	82.29±0.47	95.46±0.36	65.54±0.66	88.76±0.42
DTL+Aug(Ours)	MobileNetV2	**84.63±0.51**	**95.29±0.27**	**67.94±0.40**	**88.49±0.25**

Note: The results of the SOTA models are taken from the SCAN [31].

The "+Aug" suffix indicates the application of batch-level data augmentation techniques such as MixUp, CutMix, or ResizeMix during fine-tuning. The consolidated results are presented in Tables I and II. Table I summarizes the performances of backbone models without fine-tuning, while Table II presents the results of models fine-tuned under different training strategies, including FEL, FETL, DTL, DTLS, and baseline DL. Hence both the inherent capability of pretrained encoders and the improvements obtained through fine-tuning processes under few-shot learning settings is obtained.

Moreover, we executed additional experiments based on the parameters specified in Tables III, IV, and V. These experiments are designed to compare our model's performance against benchmarks set in prior research, such as SCAN [31], MetaMed [30], and PFEMed [37], using the datasets referred to in Section [12]–[14]. 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. In addition to dataset characteristics, model capacity in terms of parameter count and training computation cost (FLOPs) plays a significant role for few-shot classification performance. As detailed in Table I, the tested models vary widely in size, ranging from lightweight architectures such as MobileNetV2 (3.5M parameters, 0.32G FLOPs) to heavyweight models like ViT-Base (86.6M parameters, 17.6G FLOPs). In our experiments, while compact models such as MobileNetV2 and DenseNet121 generally provided stable performance across few-shot setups due to their simplicity and lower risk of overfitting, larger models like ViT-Base demonstrated notably strong performance when appropriately fine-tuned on Derm7pt. These findings highlight that smaller capacity model alone is not a disadvantage in few-shot skin disease classification; rather, the effectiveness of large models depends on the availability of diverse and well-aligned data for adaptation.

Following the evaluation of Table I, several backbone architectures, including MobileNetV2, DenseNet121, ViT-Base, and ResNet50, were selected for fine-tuning experiments due to their consistent performance across different datasets and shot configurations. These models demonstrated a favorable balance between model complexity and few-shot generalization capability. To further investigate the impact of contrastive self-supervised pretraining, we included ResNet50 models pretrained with methods such as SimCLR, SimSiam, and MoCo-v3 in our evaluations.

We have also included foundation models such as BioViL and MedCLIP in our evaluation by utilizing their vision encoders directly for few-shot classification. These models were originally pretrained on chest X-ray datasets using paired image–text data. When applied to our skin disease datasets, they performed below ImageNet-pretrained counterparts. This is likely due to a domain mismatch, as these models were trained on radiographic images with associated clinical text, whereas our datasets consist of dermoscopic and clinical skin images without accompanying text. Additionally, since we did not perform further fine-tuning due to lack of domain-specific textual data, their representations could not be adapted to the target domain, limiting their effectiveness in few-shot settings.

First, across all datasets, FETL models consistently outperformed FEL models, particularly in lower-shot settings (e.g., 2W1S, 5W1S), highlighting the critical advantage of using pretrained ImageNet weights in few-shot scenarios. As the number of shots increased, transfer learning-based models, especially DTL and DTL, demonstrated significant gains over both FEL and FETL models. For instance, in the SD-198 dataset under the 5W5S setup, MobileNetV2 achieved 83.22% accuracy with FETL, while DTL with augmentation (DTL+Aug) reached 88.49%, clearly illustrating the impact of full fine-tuning and advanced data augmentation strategies. Similarly, DenseNet121 models showed a comparable trend: 84.02% with FETL and 88.76% with DTL+Aug. The behaviour of the models are similar in Derm7pt dataset. 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 tolerate these diversities. As the number of shots increases, models trained non-episodically become more successful in classifications.

When observing Derm7pt results, ViT-Base models adapted with LoRA fine-tuning exhibited outstanding performance, achieving over 93% accuracy in low-shot settings (2W5S and 5W5S). This suggests that large-capacity models, when fine-tuned properly on diverse datasets, can effectively overcome the challenges posed by few-shot classification.

The evaluation results presented in Table II demonstrate that among the contrastive pretraining strategies tested, SimCLR consistently yields the most favorable performance across all datasets when combined with augmentation and supervised fine-tuning (DTLS). For example, in the Derm7pt dataset, ResNet50(SimCLR)+Aug outperforms SimSiam and MoCo-v3 in nearly all configurations, particularly in 2W5S and 5W5S settings, where it achieves 80.58% and 63.26% accuracy, respectively. On ISIC2018, which has fewer classes and a more homogeneous modality, all contrastive methods perform comparably, but SimCLR still maintains a slight edge. These results suggest that SimCLR-based self-supervised pretraining provides a more transferable representation for downstream few-shot tasks, especially when followed by targeted fine-tuning with data augmentation. Overall, the choice of contrastive method can significantly affect performance, highlighting the importance of selecting an appropriate pretraining strategy for different dataset characteristics.

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 2W1S with FETL and 64.83% with DTL, and 68.82% for 2W10S with

TABLE IV
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 [11]	Conv6	64.50	73.50	79.70	-	-	-
MetaMed - Transf. Learn. [30]	Conv4	66.88	73.88	81.37	54.83	59.33	69.75
MetaMed - Normal Aug. [30]		72.75	75.62	81.37	54.83	59.33	69.75
MetaMed - CutOut [30]		70.37	77.62	81.87	55.50	65.41	69.75
MetaMed - MixUp [30]		75.37	78.25	84.25	58.50	61.25	71.00
MetaMed - CutMix [30]		73.25	76.87	80.62	58.66	61.50	66.50
PT-MAP [74]	WRN	68.15	70.87	74.19	53.17	55.61	59.57
Baseline+ [73]		64.77	70.27	74.67	53.20	54.16	57.87
NegMargin [80]		71.33	72.67	75.17	60.69	57.58	63.04
Baseline [73]		68.77	71.03	76.97	56.80	59.20	65.22
PFEMed [37]		**81.69**	**83.87**	85.14	**66.94**	69.78	73.81
DTL+Aug (Ours)	ResNet50	73.57	78.28	83.47	59.79	65.86	72.70
DTL+Aug (Ours)	DenseNet121	62.42	65.06	67.85	45.74	48.66	52.34
DTL+Aug (Ours)	ViT(Base)	62.85	65.64	67.98	47.22	49.55	53.25
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	86.40	66.84	71.15	76.48
DTL+ResizeMix (Ours)		78.02	82.75	86.69	65.56	70.87	**76.97**
DTL+AllAug (Ours)		78.96	83.21	**86.83**	66.12	**71.28**	76.94

Note: The results of the SOTA models are taken from the PFEMed [37].

FETL and 85.18% with DTL (Table II). Episodic learning-based methods tend to remain superficial in training.

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 ([33]) that questions the practicality of complex few-shot training methods.

Our proposed DTL approach exhibited promising performance on datasets containing clinical data, particularly SD-198. However, its performance was relatively lower on dermoscopic datasets such as ISIC2018 and, to some extent, Derm7pt. As illustrated in Table II, DenseNet121-based models generally performed worse than MobileNetV2-based counterparts on ISIC2018, especially under low-shot conditions such as 2W1S and 3W1S. For example, under the 2W1S setup, DenseNet121 DTL achieved 67.51% accuracy, while MobileNetV2 DTL reached 69.5%. This gap can be attributed to DenseNet121's tendency to overfit in low-data regimes. Moreover, ISIC2018's structure (with only four base and three novel classes) offers limited variation for deeper architectures to generalize effectively. On Derm7pt, both MobileNetV2 and DenseNet121 showed more comparable results, with DenseNet121+Aug performing competitively in 5W5S. Nonetheless, across both datasets, neither model outperformed recent state-of-the-art methods such as SCAN or PFEMed. These results emphasize the need for improved techniques that specifically address the unique challenges of dermoscopic image classification, including fine-grained visual patterns and modality-specific artifacts. Notably, the ViT(Base)+LoRA+Aug model achieved the highest accuracy across several configurations on Derm7pt (e.g., 2W5S and 5W5S).

Our experiments further demonstrate that the transfer learning strategy in the FETL,DTL and DTLS methods effectively incorporates advances from natural image domains into skin disease classification tasks. Datasets include images with hair, ruler marks, varying body regions, and differences in skin tone. These artifacts may introduce instability in the models. However, leveraging data augmentation techniques like Flip and Resize, along with mix-based methods such as MixUp, CutMix, and ResizeMix, not only boosted model accuracy but also enhanced model robustness. A more detailed analysis and comprehensive discussion of these findings can be found in Section VI.

One limitation of our study is the absence of evaluation on independent clinical datasets, due to limited access beyond public benchmarks. While our models generalize well on standard datasets, future work will focus on external clinical validation to better assess real-world applicability.

C. Model Explainability

To improve interpretability, we applied three common explainability techniques: prototype visualization with t-SNE [82], Grad-CAM [83], and predictive uncertainty estimation using Monte Carlo dropout [84]. All results were obtained using the DTL+Aug MobileNetV2 model across all datasets.

Grad-CAM maps (Figure 6), which are obtained from the last convolutional layer of the embedding extractor, show that the model generally attends to lesion regions in clinical datasets (SD-198 and Derm7pt), while attention in dermoscopic images (ISIC2018) is more scattered, which helps explain some misclassifications. Moreover, the t-SNE visualization of 20 randomly selected SD-198 classes in Figure 6(d)

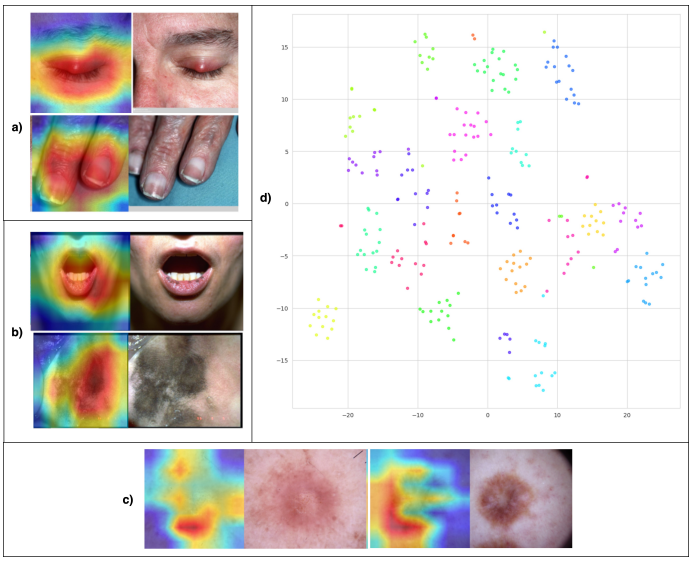

Fig. 6. (a) Grad-CAM on SD-198, (b) Grad-CAM on Derm7pt, (c) Grad-CAM on ISIC2018, (d) t-SNE of 20 randomly selected SD-198 novel classes.

demonstrates that the base-trained model, despite never observing novel classes, is still able to form reasonably separable clusters in the embedding space.

In addition, we performed t-SNE visualization during testing using prototypical embeddings, particularly focusing on 2-way and 3-way episodes. Across all three datasets, distinct clustering is observed in the embedding space; however, clustering patterns are clearer in the clinical datasets (SD-198 and Derm7pt), while ISIC2018 remains less separable due to its higher visual similarity among classes (Figure 7).

Finally, the uncertainty analysis (Figure 8) revealed complementary insights at two levels. At the episode level, predictive entropy showed a negative correlation with accuracy (Pearson $r = -0.49$ for 2-way 1-shot and $r = -0.64$ for 2-way 5-shot). At the query level, misclassified samples tended to have higher entropy values, indicating that uncertainty is a useful signal for identifying unreliable predictions. Increasing the number of support samples from 1 to 5 reduced overall uncertainty and strengthened its correlation with accuracy, showing that predictions became more stable and entropy served as a more reliable error indicator. In addition to the overall decrease, entropy distributions became narrower with more support samples, indicating more stable predictions across episodes.

D. 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 [31] and PFEMed [37] 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 evaluation on the SD-198 dataset, we enhanced our Deep Transfer Learning (DTL) models using various data augmentation techniques and compared them against recent state-of-the-art methods, including SCAN, EASY, PT+NCM, and others (Table III). Unlike the SCAN model, which uses a WRN-28-10 backbone with an unsupervised clustering branch, our models leverage lighter backbones such as MobileNetV2 and ViT, combined with ImageNet pretraining and multiple augmentation strategies (e.g., MixUp, CutMix, ResizeMix). Notably, our MobileNetV2-based DTL+AllAug model outperformed all prior methods in every configuration, achieving the highest F1-scores across all 2-Way and 5-Way shot settings. This indicates that even with a simpler architecture and traditional transfer learning, effective data augmentation can result in state-of-the-art performance.

On the ISIC2018 dataset, we compared our results with benchmarks established by PFEMed, MetaMed, PT-MAP, and others (Table IV). PFEMed, which incorporates a dual-encoder architecture with variational modeling, achieved strong results, particularly in low-shot 3-Way setups. However, our MobileNetV2 DTL+AllAug model achieved the highest accuracy in several 2-Way and 3-Way configurations, including 2W10S (86.83%) and 3W10S (76.94%). Furthermore, our ViT-Base-based model also showed competitive results, confirming that transformer backbones can generalize well in few-shot medical image settings when properly fine-tuned. These findings suggest that our approach can match or even surpass more complex meta-learning-based methods, particularly when sufficient training shots are available.

We further compared our models against state-of-the-art methods on the Derm7pt dataset. As shown in Table V, our ViT(Base)+Aug DTL model achieved the highest accuracies in both 2-Way 1-Shot (86.35%) and 2-Way 5-Shot (93.68%) settings, outperforming strong baselines such as SCAN and PFEMed. In addition, MobileNetV2 and DenseNet121-based DTL models achieved competitive performance, especially when paired with augmentation techniques such as MixUp and CutMix. These results reinforce the notion that a well-designed transfer learning framework, which combines pre-trained weights, architectural diversity, and batch-level augmentation, can effectively tackle the unique challenges posed by long-tailed dermatological datasets, without relying on meta-learning or external pretraining data.

VI. Additional Analysis

We conducted additional analysis to examine how various augmentation strategies affect the performance of our DTL models on three benchmark datasets: SD-198, Derm7pt, and ISIC2018. We employed the following strategies: (1) Base (random horizontal and vertical flips at 45°), (2) CutMix, (3) MixUp, (4) ResizeMix, and (5) AllAug, which combines all the aforementioned methods on top of the Base augmentation. Detailed results for the SD-198, Derm7pt, and ISIC2018 datasets are presented in Tables VI.

We previously mentioned in Section V-B that artifacts in skin disease datasets compromise model stability, yet mix-based augmentation techniques effectively address these issues. Notably, the SD-198 dataset contains clinical images with varying artifacts (i.e. high scale variances within classes)

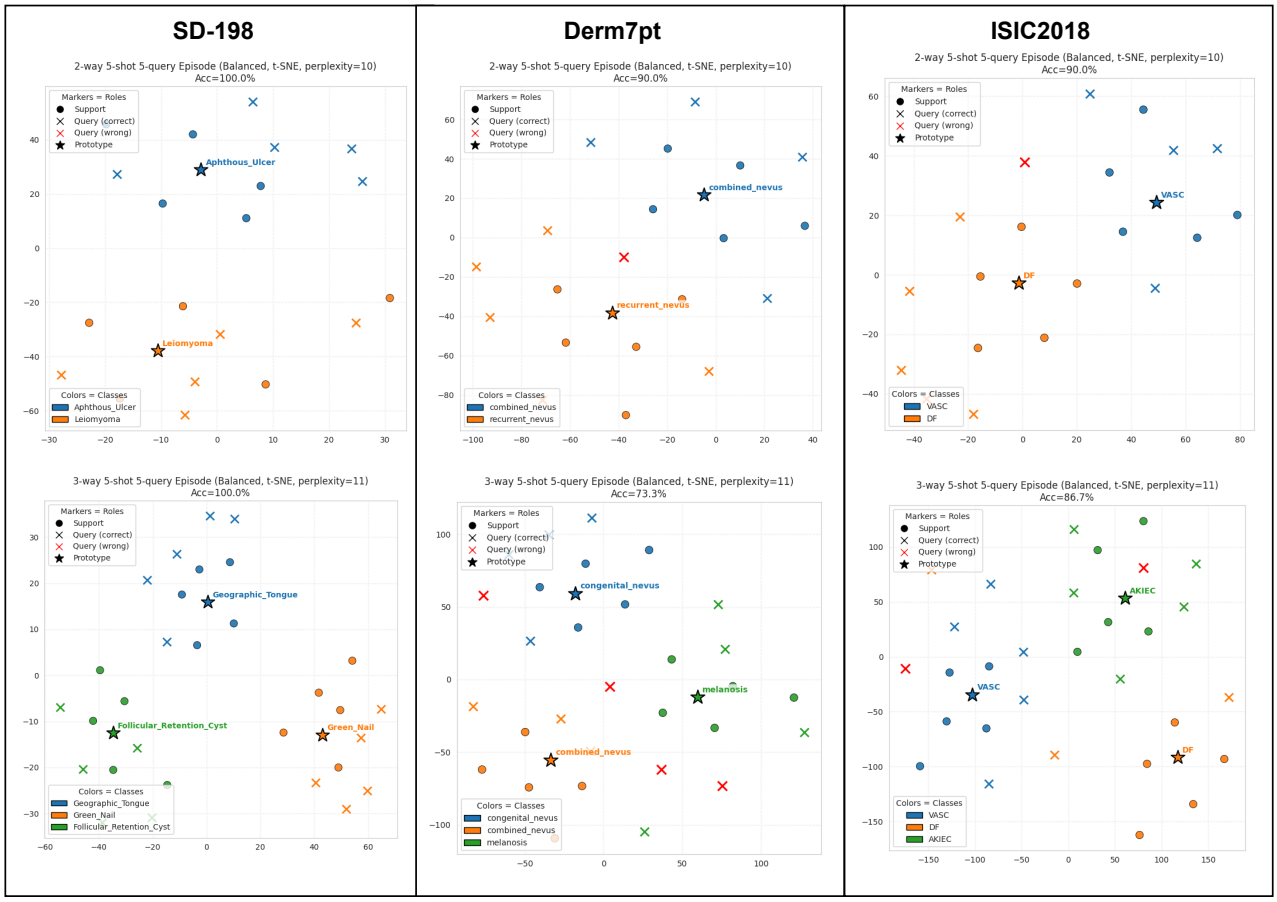

Fig. 7. Prototype visualization with t-SNE on ISIC2018, Derm7Pt, and SD-198. Please refer to the digital version for better visibility.

TABLE V
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 [26]	Conv4	59.98±1.28	70.62±1.3
SCAN [31]		61.42±1.49	72.58±1.28
Meta-DermDiagnosis [11]	Conv6	61.8	76.9
SCAN [31]		62.80±1.34	76.65±1.21
NCA [79]	WRN-28-10	56.32±1.29	67.18±1.15
Baseline [73]		59.43±1.34	74.28±1.14
S2M2_R [72]		61.37±1.33	79.83±1.34
NegMargin [80]		58.00±1.44	70.12±1.30
PT+NCM [74]		60.92±1.68	74.33±1.48
PEMb E_NCM [79]		60.40±1.72	72.63±1.48
EASY [81]		61.02±1.67	75.98±1.41
SCAN [31]		66.75±1.35	82.57±1.13
PFEMed [37]		71.15	80.27
DTL+Aug (Ours)	ResNet50	61.62±0.53	78.86±0.47
DTLS+Aug (Ours)	ResNet50(SimCLR)	63.91±0.38	80.58±0.44
DTL+Aug (Ours)	MobileNetV2	60.37±0.69	77.41±0.41
DTL+Aug (Ours)	DenseNet121	62.56±0.40	81.00±0.39
DL (Ours)	MoCo-v3(ViT)	**88.40±0.55**	92.18±0.63
DTL+Aug (Ours)	ViT-Base	86.35±0.66	**93.68±0.48**

Note: The results of the SOTA models are taken from the SCAN [31].

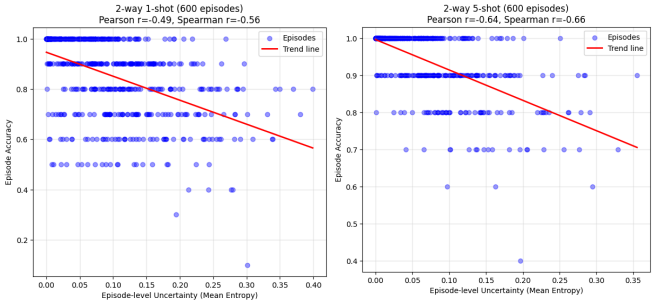

Fig. 8. Episode-level predictive uncertainty (mean entropy) versus episode accuracy for the SD-198 dataset with the MobileNetV2 model using Monte Carlo dropout.

due to different camera setups. Since ResizeMix resizes one image and overlays it onto another, it enhances the model's resilience to such variations. Consequently, ResizeMix outperformed CutMix, MixUp, and AllAug by approximately 3% on the SD-198 dataset.

In contrast, dermoscopy datasets, such as Derm7pt and ISIC2018, typically involve specialized imaging devices that capture deeper layers of the skin at uniform depths, minimizing distance-related variability and other artifacts. As a result,

TABLE VI
PERFORMANCE COMPARISON OF MODELS AND AUGMENTATION TECHNIQUES ON **SD-198, DERM7PT, ISIC2018** DATASETS, REPORTING ACCURACIES(%) METRICS.

Dataset Model	Method	2W-1S	2W-5S	5W-1S	5W-5S
SD-198 MobileNetV2	Base	82.42	94.40	64.06	86.67
	Base+CutMix	83.00	94.77	65.48	87.63
	Base+Mixup	82.29	94.48	64.56	87.22
	Base+ResizeMix	**84.63**	**95.29**	**67.94**	**88.49**
	Base+AllAug	83.31	95.02	65.99	88.04
SD-198 DenseNet121	Base	81.21	94.56	62.61	86.95
	Base+CutMix	77.61	93.95	58.23	86.44
	Base+Mixup	78.88	94.40	61.21	87.10
	Base+ResizeMix	**82.29**	**95.46**	**65.54**	**88.76**
	Base+AllAug	80.80	95.27	63.59	88.72
Derm7pt MobileNetV2	Base	59.65	75.41	31.82	56.10
	Base+CutMix	**60.97**	77.06	**33.84**	59.02
	Base+Mixup	60.65	76.92	33.21	58.57
	Base+ResizeMix	60.37	**77.41**	33.54	**59.45**
	Base+AllAug	60.53	77.10	33.37	59.05
Derm7pt DenseNet121	Base	60.00	77.25	32.04	57.68
	Base+CutMix	61.57	78.65	**35.46**	60.58
	Base+Mixup	**62.56**	**81.00**	34.75	**61.91**
	Base+ResizeMix	61.09	79.18	33.89	61.22
	Base+AllAug	60.88	78.30	33.19	59.95
Dataset Model	**Method**	**2W-1S**	**2W-5S**	**3W-1S**	**3W-5S**
ISIC2018 MobileNetV2	Base	64.83	81.63	49.51	69.35
	Base+CutMix	64.94	81.79	49.93	69.86
	Base+Mixup	**66.34**	82.95	**51.66**	71.15
	Base+ResizeMix	64.99	82.75	50.34	70,87
	Base+AllAug	65.55	**83.21**	50.40	**71.28**
ISIC2018 DenseNet121	Base	55.55	64.69	38.63	46.47
	Base+CutMix	54.45	60.78	38.63	44.32
	Base+Mixup	55.20	64.42	39.72	**48.98**
	Base+ResizeMix	55.90	63.57	**39.85**	48.27
	Base+AllAug	**56.67**	**65.06**	39.81	48.66

ResizeMix demonstrated similar performance to CutMix and MixUp on these two datasets.

Finally, we investigated how the architectural characteristics of MobileNetV2 and DenseNet121 influence the outcomes of these augmentations. Both backbones achieved comparable performance on the SD-198 and Derm7pt datasets. However, as discussed in the Discussion section, DenseNet121 suffered from overfitting on the ISIC2018 dataset and consequently failed to deliver robust results.

VII. CONCLUSION AND FUTURE WORK

In this study, we explored the effectiveness of supervised and self-supervised transfer learning strategies combined with few-shot learning to classify rare skin diseases under long-tail data distributions. While prior work has highlighted the challenges of imbalance and data scarcity in this domain, there remains limited research evaluating the combined impact of transfer learning and few-shot methods in a unified setting. To address this gap, we systematically evaluated five training paradigms (FETL, FEL, DTL, DTLS, and DL) across three benchmark skin image datasets. Our findings demonstrate that DTL models, particularly those using MobileNetV2 backbones and enhanced with MixUp, CutMix, and ResizeMix, consistently outperformed alternative approaches. In addition, transformer-based models such as ViT-Base, fine-tuned via LoRA, achieved state-of-the-art performance on the Derm7pt dataset.

These results reaffirm the strength of conventional transfer learning in few-shot medical imaging tasks and highlight the value of appropriate data augmentation and backbone selection. While episodic and self-supervised methods showed advantages in specific low-shot scenarios, their performance was generally less stable compared to supervised DTL approaches. Additionally, contrastive self-supervised pretraining strategies such as SimCLR and MoCo-v3 were also applied and yielded competitive results in selected scenarios. The datasets used in this study only contained class labels and did not provide additional clinical context such as lesion descriptions, anatomical site annotations, or physician notes. For this reason, vision–language foundation models (e.g., BioViL, MedCLIP) were evaluated in a zero-shot setting and showed limited transferability due to domain mismatch and the absence of text-based adaptation.

Future research may benefit from exploring domain-specific self-supervised pretraining strategies, lightweight transformer fine-tuning methods, and the integration of multimodal approaches. Recently introduced dermatology datasets with rich textual annotations, such as DermaCon-IN [85], indicate that the field is moving toward cross-modal resources. Building upon these, future work could explore integration with CLIP-based vision–language models and multimodal LLMs such as GPT-4o or BiomedCLIP to combine lesion images with textual or patient-level context, enabling interactive diagnostic support and zero-shot generalization. Finally, extending the framework to other medical imaging modalities (e.g., X-rays, histopathology, fundus imaging) in parallel with multimodal integration could further support robust classification systems for low-data and rare disease applications.

REFERENCES

[1] D. Chen, Y. Bai, W. Shen, Q. Li, L. Yu, and Y. 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*, 2023, pp. 23 869–23 878.

[2] B. Wang, Q. Li, and Z. You, "Self-supervised learning based transformer and convolution hybrid network for one-shot organ segmentation," *Neurocomputing*, vol. 527, pp. 1–12, 2023.

[3] Q. Hu, Y. Chen, J. Xiao, S. Sun, J. Chen, A. L. Yuille, and Z. Zhou, "Label-free liver tumor segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7422–7432.

[4] I. Mazumdar and J. Mukherjee, "Fully automatic mri brain tumor segmentation using efficient spatial attention convolutional networks with composite loss," *Neurocomputing*, vol. 500, pp. 243–254, 2022.

[5] S. Mishra, Y. Zhang, L. Zhang, T. Zhang, X. S. Hu, and D. 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*. Springer, 2022, pp. 721–731.

[6] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, 2023.

[7] B. H. Y. Chung, J. F. T. Chau, and G. K.-S. Wong, "Rare versus common diseases: a false dichotomy in precision medicine," *NPJ Genomic Medicine*, vol. 6, no. 1, p. 19, 2021.

[8] K. Liopyris, S. Gregoriou, J. Dias, and A. J. Stratigos, "Artificial intelligence in dermatology: challenges and perspectives," *Dermatology and Therapy*, vol. 12, no. 12, pp. 2637–2651, 2022.

[9] M. K. Hasan, M. A. Ahamad, C. H. Yap, and G. Yang, "A survey, review, and future trends of skin lesion segmentation and classification," *Computers in Biology and Medicine*, p. 106624, 2023.

[10] P. Yao, S. Shen, M. Xu, P. Liu, F. Zhang, J. Xing, P. Shao, B. Kaffenberger, and R. X. Xu, "Single model deep learning on imbalanced small datasets for skin lesion classification," *IEEE transactions on medical imaging*, vol. 41, no. 5, pp. 1242–1254, 2021.

[11] K. Mahajan, M. Sharma, and L. 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*, 2020, pp. 730–731.

[12] X. Sun, J. Yang, M. Sun, and K. 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*. Springer, 2016, pp. 206–222.

[13] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[14] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. 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.

[15] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and meta: A new perspective on few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 669–10 686, 2023.

[16] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4650–4666, 2022.

[17] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Retain and recover: Delving into information loss for few-shot segmentation," *IEEE Transactions on Image Processing*, 2023.

[18] G. Cheng, L. Cai, C. Lang, X. Yao, J. Chen, L. Guo, and J. Han, "Spnet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[19] H. Quan, X. Li, D. Hu, T. Nan, and X. Cui, "Dual-channel prototype network for few-shot pathology image classification," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[20] H. Jiang, M. Gao, H. Li, R. Jin, H. Miao, and J. Liu, "Multi-learner based deep meta-learning for few-shot medical image classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 17–28, 2022.

[21] L. Sun, M. Zhang, B. Wang, and P. Tiwari, "Few-shot class-incremental learning for medical time series classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 1872–1882, 2023.

[22] X. Li, Y. Tan, B. Liang, B. Pu, J. Yang, L. Zhao, Y. Kong, L. Yang, R. Zhang, H. Li *et al.*, "Tkr-fsod: Fetal anatomical structure few-shot detection utilizing topological knowledge reasoning," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[23] Y. He, T. Li, R. Ge, J. Yang, Y. Kong, J. Zhu, H. Shu, G. Yang, and S. Li, "Few-shot learning for deformable medical image registration with perception-correspondence decoupling and reverse teaching," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1177–1187, 2021.

[24] M. K. Hasan, M. T. E. Elahi, M. A. Alam, M. T. Jawad, and R. Martí, "Dermoexpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation," *Informatics in Medicine Unlocked*, vol. 28, p. 100819, 2022.

[25] S. Jain, U. Singhania, B. Tripathy, E. A. Nasr, M. K. Aboudaif, and A. K. Kamrani, "Deep learning-based transfer learning for classification of skin cancer," *Sensors*, vol. 21, no. 23, p. 8142, 2021.

[26] V. Prabhu, A. Kannan, M. Ravuri, M. Chaplain, D. Sontag, and X. Amatriain, "Few-shot learning for dermatological disease diagnosis," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 532–552.

[27] X. Li, L. Yu, Y. Jin, C.-W. Fu, L. Xing, and P.-A. 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*. Springer, 2020, pp. 357–366.

[28] D. Zhang, M. Jin, and P. Cao, "St-metadiagnosis: Meta learning with spatial transform for rare skin disease diagnosis," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 2153–2160.

[29] W. Zhu, H. Liao, W. Li, W. Li, and J. 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*. Springer, 2020, pp. 330–339.

[30] R. Singh, V. Bharti, V. Purohit, A. Kumar, A. K. Singh, and S. K. Singh, "Metamed: Few-shot medical image classification using gradient-based meta-learning," *Pattern Recognition*, vol. 120, p. 108111, 2021.

[31] S. Li, X. Li, X. Xu, and K.-T. Cheng, "Dynamic subcluster-aware network for few-shot skin disease classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[32] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. 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*. Springer, 2020, pp. 124–141.

[33] S. Laenen and L. Bertinetto, "On episodes, prototypical networks, and few-shot learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 581–24 592, 2021.

[34] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. 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*. Springer, 2020, pp. 266–282.

[35] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 403–412.

[36] J. Xiao, H. Xu, D. Fang, C. Cheng, and H. Gao, "Boosting and rectifying few-shot learning prototype network for skin lesion classification based on the internet of medical things," *Wireless Networks*, vol. 29, no. 4, pp. 1507–1521, 2023.

[37] Z. Dai, J. Yi, L. Yan, Q. Xu, L. Hu, Q. Zhang, J. Li, and G. Wang, "Pfemed: Few-shot medical image classification using prior guided feature enhancement," *Pattern Recognition*, vol. 134, p. 109108, 2023.

[38] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[42] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[43] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[45] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9640–9649.

[46] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[47] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[48] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[49] Y. Liu, S. Zhang, J. Chen, Z. Yu, K. Chen, and D. Lin, "Improving pixel-based mim by reducing wasted modeling capability," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5361–5372.

[50] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle *et al.*, "Making the most of text semantics to improve biomedical vision–language processing," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[51] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2022, 2022, p. 3876.

[52] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 201–208.

[53] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[54] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7310–7311.

[55] W. Ying, Y. Zhang, J. Huang, and Q. Yang, "Transfer learning via learning to transfer," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5085–5094.

[56] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3712–3722.

[57] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification," *Computer methods and programs in biomedicine*, vol. 193, p. 105475, 2020.

[58] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, "A gan-based image synthesis method for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105568, 2020.

[59] L. Liu, L. Mou, X. X. Zhu, and M. Mandal, "Automatic skin lesion classification based on mid-level feature learning," *Computerized Medical Imaging and Graphics*, vol. 84, p. 101765, 2020.

[60] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 19–29, 2019.

[61] S. Atasever, N. Azginoglu, D. S. Terzi, and R. Terzi, "A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning," *Clinical Imaging*, vol. 94, pp. 18–41, 2023.

[62] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[63] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, vol. 2, no. 3, p. 4, 2018.

[64] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," *arXiv preprint arXiv:1807.05960*, 2018.

[65] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[66] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[67] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[68] X. Li, X. Yang, Z. Ma, and J.-H. Xue, "Deep metric learning for few-shot image classification: A review of recent developments," *Pattern Recognition*, p. 109381, 2023.

[69] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3018–3027.

[70] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[71] R. Kumar, T. Deleu, and Y. Bengio, "The effect of diversity in meta-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 8396–8404.

[72] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, and V. 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*, 2020, pp. 2218–2227.

[73] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *arXiv preprint arXiv:1904.04232*, 2019.

[74] Y. Hu, V. Gripon, and S. Pateux, "Leveraging the feature distribution in transfer-based few-shot learning," in *International Conference on Artificial Neural Networks*. Springer, 2021, pp. 487–499.

[75] W. Zhu, W. Li, H. Liao, and J. Luo, "Temperature network for few-shot learning with distribution-aware large-margin metric," *Pattern Recognition*, vol. 112, p. 107797, 2021.

[76] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[77] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[78] J. Qin, J. Fang, Q. Zhang, W. Liu, X. Wang, and X. Wang, "Resizemix: Mixing data with preserved object information and true labels," *arXiv preprint arXiv:2012.11101*, 2020.

[79] Z. Wu, A. A. Efros, and S. X. Yu, "Improving generalization via scalable neighborhood component analysis," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 685–701.

[80] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. 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*. Springer, 2020, pp. 438–455.

[81] Y. Bendou, Y. Hu, R. Lafargue, G. Lioi, B. Pasdeloup, S. Pateux, and V. Gripon, "Easy—ensemble augmented-shot-y-shaped learning: State-of-the-art few-shot classification with simple components," *Journal of Imaging*, vol. 8, no. 7, p. 179, 2022.

[82] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[83] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[84] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[85] S. S. Madarkar, M. Madarkar, T. Prakash, K. R. Mopuri, V. MV, K. Sathwika, A. Kasturi, G. D. Raj, P. Supranitha, H. Udai *et al.*, "Dermaconin: A multi-concept annotated dermatological image dataset of indian skin disorders for clinical ai research," *arXiv preprint arXiv:2506.06099*, 2025.

Zeynep ÖZDEMİR received her B.S. and M.S. degrees in Computer Engineering from Ankara University, Turkey, in 2016 and 2018, respectively. She has been a Research Assistant in the Department of Computer Engineering at Ankara University since 2018. Her research interests include machine learning and deep learning applications in the medical domain, with a focus on computer vision, medical data analysis, and rare disease classification. During her M.S. studies, she worked on GPS, GPRS, and GSM-based real-time map tracking systems. She has also been an active member of the RISE-MICCAI reading group for the past year, further expanding her expertise in medical imaging and computational methods.

Dr. Hacer Yalim Keles completed her B.Sc., M.Sc., and Ph.D. in Computer Engineering at Middle East Technical University, Turkey, in 2002, 2005, and 2010, respectively. Her Ph.D. thesis was honored with the Thesis of the Year award by the Prof. Dr. Mustafa Parlar Education and Research Foundation in 2010. Between 2000 and 2007, she contributed as a researcher at The Scientific and Technological Research Council of Turkey (TUBITAK), focusing on pattern recognition using multimedia data, including audio and video.

In 2010, she founded her own R&D company with a grant from the Ministry of Industry and Trade of Turkey. Her project SOYA, funded by TUBITAK in 2011, was later recognized as one of the best venture projects, leading to an opportunity in Silicon Valley for potential investments. Dr. Keles was an Assistant Professor at the Department of Computer Engineering, Ankara University, from 2013 to 2021, and is currently an Associate Professor at the Department of Computer Engineering, Hacettepe University.

Her research primarily spans computer vision and machine learning, with a focus on learning algorithms for limited data and deep generative models. She has contributed to sign and gesture recognition, generative adversarial networks, image inpainting, and image segmentation domains. Moreover, she collaborates on diverse projects involving aerial and medical images, speech signals, textual, geophysical, and hyperspectral data analysis with her graduate students.

Dr. Ömer Özgür TANRIÖVER received the B.Sc. degree in computer engineering, the M.Sc. and Ph.D. degrees in information systems from Middle East Technical University, Ankara, Turkey. Previously, he has served as a Certified Information Systems Auditor (CISA) with the Information Management Department, Banking Regulation Agency of Turkey. Currently, he is with Computer Engineering Department of Ankara University as an Associate Professor. His current research interests include applications of machine learning in human computer interaction, information systems security and software process.