Meta-Learning for Zero-Shot Skin Lesion Classification Across Unseen Smartphone Devices

by

Ishmam Azmine
24141206
MD Fuyad Ibnay Rafi
22141018
Shadab Uddin Dewan
23101555
Quazi Tousif Ishraq
22101228

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2026

© 2026. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ishmam Azmine

24141206

Fuad

MD Fuyad Ibnay Rafi

22141018

Shadab Uddin Dewan

23101555

Tousif

Quazi Tousif Ishraq

22101228

Approval

The thesis/project titled "Meta-Learning for Zero-Shot Skin Lesion Classification Across Unseen Smartphone Devices" submitted by

1. Ishmam Azmine (24141206)
2. MD Fuyad Ibnay Rafi (22141018)
3. Shadab Uddin Dewan (23101555)
4. Quazi Tousif Ishraq (22101228)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 31, 2026.

Examining Committee:

Supervisor:
(Member)

Dr. Chowdhury Mofizur Rahman

Professor
Department of Computer Science and Engineering
Institution

Co-Supervisor:
(Member)

MD. Sabbir Ahmed

Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The biggest challenge to automated classification of skin lesions is robust cross-domain generalization especially when models trained on dermoscopic images are applied to smartphone images. This paper assesses how meta-learning can achieve cross-domain robustness on a strict zero-shot protocol, where no target-domain images, labels, or hyperparameter feedback are used during training or model selection. Models are only trained on HAM10000, BCN20000 dermoscopy datasets and evaluated on PAD-UFES-20 smartphone dataset with a single six-class taxonomy. We compare the baselines of supervised transfer learning with several meta-learning paradigms, which include Prototypical Networks, Meta-Baseline, FEAT, MetaOptNet. Experiments are evaluated on convolutional and transformer-based backbones and evaluated in a similar episodic format. Macro-averaged F1-score is used as the main measure of performance because it accounts for class imbalance. Across backbones, the monitored models record a significant drop in source to target performance irrespective of good source domain validation outcomes. Meta-learning techniques mitigate this degradation, with relative macro-F1 gains of about 20–70% more than supervised baselines in a variety of backbone architectures in various configurations. These findings prove that episodic meta-learning can reduce the domain-induced failures in the task of classifying skin lesions in the absence of any target-domain exposure. The suggested evaluation environment captures the actual restrictions of teledermatology implementation and gives empirical data of meta-learning as a potential approach towards zero-shot cross-domain medical imaging analysis.

Keywords: Meta-Learning, Domain Generalization, Zero-Shot Classification, Skin Lesion Analysis, Cross-Device Robustness

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis has been completed without any major interruption.

Secondly, to our supervisor Dr. Chowdhury Mofizur Rahman sir and our co-supervisor MD. Sabbir Ahmed sir for their kind support and advice in our work. They helped us whenever we needed help.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Chapter 1

Introduction

1.1 Background

Automated classification of skin lesions by means of deep learning has rapidly matured over the past decade [29]. Large, curated dermoscopic datasets and state-of-the-art convolutional and transformer backbones have unlocked classifiers with impressive accuracy when tested on held out images from the same acquisition hardware and clinical workflow [24][29]. These systems are trained to look for visual patterns - networks of pigments, asymmetry, issues along borders, texture - that correlate to malignancy, or benign diagnoses [29]. Yet the research-to-deployment gap is large because real-world images one encounters on the street are much more diverse than those used in developing the model [34][33]. In particular, there is a strong and systematic device effect: Photos taken with different cameras, sensors, lenses, illumination or compression pipelines can create images that look totally different even if the underlying lesion is exactly the same [34][33]. Consumer smartphones bring to the table many of these variations. They vary between manufacturers and models for colour response, white balance, noise on a pixel by pixel level, strength of the jpg compression, lens artifacts, and default image processing pipelines [34][24]. Additionally, user behavior (distance, angle, focus, illuminate flash) also adds to the inter-device variability [33]. As a result, models trained on clinical dermatoscope images - which tend to be higher resolution and controlled (and consistent) - often exhibit well-over 50 percent degradation in performance when used to reverse the roles and apply to smartphone images taken by patients or general practitioners [34][24]. This mismatch isn't just a problem from an academic perspective: it has direct implications for the clinical utility and safety of automated screening tools designed for teledermatology and (population-based) screening [34][23]. Therefore, a strong and practical diagnostic model should have to generalize over imaging devices that it has never seen at development time [12][10].

1.2 Rational of the Study or Motivation

The driving force of this thesis is practical and clinical: smartphone photography is the most scalable and available modality for widespread screening and telemedicine of large skin areas for possible skin health concerns [34][24]. "Because many skin diseases present similarly even with different dermoscopes, if an algorithm can work well only with images of a particular dermoscope, its effectiveness for promoting early detection and expanding access is considerably reduced" [34]. At the same time, such a process has a methodological rationale: recent work in meta-learning suggests that training policies that learn to learn (i.e., training on many small tasks/episodes) can produce representations that generalize better to new tasks and distributions [12]. Episodic meta-learning leads to learning discriminative representations that generalize across task variation [12][10]. If variability added to the task consists of device-to-device variation, meta-training may put more weight on the lesion-intrinsic features while putting less weight on device artifacts [31][33]. We believe that episodic meta-learning in combination with augmentations that represent smartphone artifacts (color shifts, compression, blur, stylistic variations) is an interesting approach [33][31]. Instead of gathering labeled images from every phone model, we can design for robustness towards device variability [25][31]. This thesis therefore investigates if such an approach — a reproducible, strictly zero-shot evaluation protocol with dermatoscopic sources for training and smartphone images as an unseen test domain — will significantly lessen the device gap and lead to clinically reproducible models [31][10].

1.3 Problem Statement

Deep models for skin lesion classification exhibit high within-domain performance while there exists a consistent and meaningful failure mode known as device-driven domain shift [33][31]. Models trained on datasets of dermoscopic images collected with pooled data often fail on images that were collected using different sensors, compressions, or lighting (which fails safety by not trusting the prediction); in turn, if trained on different types of data, models cannot be trusted to investigate new cases and cannot be assumed to be safe to apply to a new working setting [34][24]. Much previous work attempts to address the challenge of domain shift either by mixing target examples with training, targeting training on a few target images, or unsupervised adaptation assuming access to unlabeled target data [31][10]. While valuable, these approaches don't encompass a zero-shot deployment in the strictest of senses, where a model is exposed to a brand new device or capture pipeline, with no space to adapt to it [12][31]. The central research problem of this thesis is to find out whether meta-learning — in this case episodic device-aware training of prototypical and related meta-learners in combination with device-sensitive augmentations and careful episode design — can lead to representations that generalize to a true unseen smartphone-domain (PAD-UFES-20) without the need for any target-domain tuning [12][33][30][10]. The goals of the investigation are to characterize (a) the extent of which meta-learning reduces the source→target performance gap relative to a supervised baseline, (b) the constructions and augmentations of episodes that led to the biggest improvements, and (c) the way in which improvements are distributed across lesion classes including clinically important minority classes [12][33][31].

1.4 Objective

- To conduct data preprocessing and dataset integration, consolidating dermoscopic datasets (HAM10000 and BCN20000), cleaning the labels, standardizing images and mapping the samples to a single unified six class taxonomy to conduct uniform experimentation.

- To train and assess various supervised baseline models based on various backbone architectures (ResNet, ConvNeXt, and others) on shared dermoscopic images, and to study their weaknesses, when tested on out-of-band smartphone images in a zero-shot context.
- To apply Prototypical Networks to every backbone architecture and define whether it is possible to train an increase in metric-based meta-learning to advantage zero-shot robustness compared to traditional supervised training.
- To make a direct comparison of optimized Prototypical Network models and their supervised counterparts, to compare the improvements in zero-shot performance and robustness.
- To conduct a detailed analysis of the most successful configurations, scale the analysis to more meta-learning techniques used on different backbones, and to evaluate their performance through extensive zero-shot evaluation measures.

1.5 Methodology in Brief

In order to design and test the proposed zero-shot skin lesion classification system, this study is based on a modular, end-to-end development system that can be divided into five steps:

- The present experiment involves the use of the controlled experimental methodology of testing zero-shot classification of skin lesions in severe cross-domain shift, between the dermoscopic-image and the smartphone-image-acquired clinical-image. The key question is to explore the possibility of enhancing the robustness of meta-learning methods in cases when no target-domain data is provided in the training, validation, and the selection of models.
- There is a rigid separation of sources and targets that is imposed during the study. The domain source includes the dermoscopic pictures of the HAM10000 and BCN20000 datasets, and they are merged to enhance sample diversity without varying acquisition conditions. All the pictures are mapped to a common set of six classes of diagnoses to guarantee consistency of labels. The PAD-UFES-20 dataset is used as the target domain, comprising the images that were collected under uncontrolled conditions on smartphones and are represented by clinical images. Only final evaluation is conducted using target-domain images, which has to be genuine in terms of zero-shot conditions.
- The approach divides the feature representation and learning strategy so as to compare the results with control. Comparisons are made between the backbone architectures, which include convolutional neural networks (ResNet, EfficientNet-B3, ConvNeXt-Tiny) and transformer-based models (DeiT-Small and DINOv2-Small). The weights of all backbones are pretrained with the publicly available one to promote stalemate optimization.

- There are two paradigms of learning. Training of conventional supervised baselines is a mini-batch optimization based on cross-entropy loss on the data of the source domain. Parallel to that, episodic meta-learning paradigms, i.e., Prototypical Networks, Meta-Baseline, FEAT, and MetaOptNet, are trained on task-based episodic sampling based on support and query sets. The aim of this training technique is to lessen overfitting to the statistics of the source domain and induce transferable representations.

- The models are all characterized by the same preprocessing pipelines, training splits and source-domain validation procedures. Hyperparameter optimization and early termination is based purely on source-domain validation. In evaluation, the models are simply applied to the target domain without fine-tuning, adaptation and calibration. The evaluation of performance is based mainly on macro-averaged F1-score, and accuracy is given as a secondary measure, focusing on the ability to be robust in the face of class imbalance and extreme domain shift.

1.6 Scopes and Challenges

The study is dedicated to the assessment of strict zero-shot generalization on the problem of skin lesion classification in case of severe cross-domain shift across devices. The models are only trained on dermoscopic images and are tested on the clinical images captured on a smartphone without any access to target-domain information during training, validation, or model selection. It is analyzed in a single-convolutional six-class diagnostics environment, and involves both convolutional and transformer-based backbone models with both supervised and meta-learning frameworks. The study does not focus on source-domain accuracy maximization but instead on robustness and transferability as they imply realistic deployment settings.

- Significant domain gap between dermoscopic and smartphone images due to differences in illumination, scale, background, and color distribution.

- Strict zero-shot constraints prohibit domain adaptation, fine-tuning, or test-time calibration.

- Computational limitations restrict extensive hyperparameter tuning and multi-seed evaluations, emphasizing controlled comparisons over statistical exhaustiveness.

- Class imbalance and label noise in skin lesion datasets increase variability in per-class performance, particularly for rare but clinically important categories.

- Limited target-domain sample size constrains detailed error analysis and increases uncertainty in performance estimates.

Nevertheless, the present research offers empirical data regarding the ability of meta-learning and device-awareness training to increase robustness on unseen smartphone images and delivers a replicable structure to future studies.

Chapter 2

Literature Review

2.1 Preliminaries

In this section, the fundamental concepts, datasets, and analysis principles defining this study are presented. It establishes the conceptual foundation that is required to understand zero-shot device generalization in skin lesion classification, applying meta-learning in episodic training paradigm,, the adopted class taxonomy, and evaluation criterion to evaluate robustness in severe domain shift. Moreover, this section describes the origins of domain shift in dermatological imaging and indicates the principles of reproducibility that will be adhered to in the course of the study.

- Zero-shot device generalization refers to the challenge in training a classification model with data obtained through one imaging modality, and evaluating it on data obtained through a completely unseen device. Without any form of adaptation, or exposure to the target domain in the course of training. In this study, models will be trained only on dermoscopic images on the HAM10000 and BCN20000 datasets [5][7] and tested only on smartphone images in the PAD-UFES-20 dataset [14]. No images from the target domain are used in the training, validation, or hyperparameter optimization. This strict separation ensures that those representations learned are not caused by target-domain characteristics and that performance reflects genuine cross-device generalization. In this context, the model must acquire device-independent features of lesions, including texture, pigmentation pattern, and structure features. Overlooking the artifacts of devices such as inconsistencies in the illumination, color imbalance, and background clutters, which have been proven to have a significant effect on clinical image analysis [4][1].

- Meta-learning, often described as “learning to learn”, will be used in this work utilizing an episodic training paradigm [3][2]. The model is not optimized on each sample, the model is trained across a distribution of classification tasks. Each training episode is modified as a small classification problem with support set and query set. The class-specific representations are created with the support set and task performance is analyzed with the query set. By repeatedly solving such episodic tasks, the model is able to learn representations that are task-generalised instead of over-fitting to one data distribution. It has been demonstrated that episodic learning is not only effective in few-shot and low-data regimes, but also in medical imaging tasks recently [13][28].

- In this type of meta-learning, the Prototypical Networks (ProtoNets)are adopted as the primary learning approach [3]. ProtoNets are a representation of each class by a prototype which is the mean embedding of the support samples of that class. The classification is carried out by classifying query samples into the closest prototype with a distance measure like the Euclidean distance or cosine distance.Prototypical Networks are selected due to their simplicity in concept and computational efficiency and interpretability in embedding space. Although ProtoNets are used as the primary baseline in the research, other meta-learning techniques such as Meta-Baseline [6], FEAT [17], and MetaOptNet [8] are subsequently studied to determine the effect of various meta-learning frameworks on zero-shot generalization.

- The article examines three datasets spanning two acquisition modalities. The training and validation are done on HAM10000 and BCN20000 datasets [5][7] in the source domain. The target domain is PAD-UFES-20 [14], which consists of clinical images gathered by smartphones and remains completely unseen during training. All diagnostic labels are mapped to a unified six-category taxonomy Actinic Keratosis (ACK), Basal Cell Carcinoma (BCC), Melanoma (MEL), Nevus (NEV), Squamous Cell Carcinoma (SCC) and Seborrheic Keratosis (SEK). Random samples of data are fixed and randomized with constant random seeds and constant CSV mappings, which ensures uniformity of results between the experiments.

- Domain shift in the classification of skin lesions is based on multiple factors including variations in camera sensors, optical characteristics, image compression, color encoding, illumination, and user handling. These may have a significant effect on the visual appearance of the same lesion across devices, resulting in distributional distortions between the source and target domains [9][11]. These changes are especially extreme in the context of transitioning between dermoscopic imaging and smartphone image. In an effort to enhance robustness to these conditions, the training is done with augmentations that are sensitive to the domain like color perturbations, compression artifacts, blurring, and style-based transformations to train to cross-device variations and to induce learning of invariant representations.

- The macro F1-score, which equally weighs all diagnostic classes, is used to estimate model performance. It is adequately applied in medical tasks with class imbalance in dermatology [4]. Other measures, such as per-class precision, recall and area under the ROC curve, are provided to give additional information. Resistance to device shift is measured with regard to the difference in the performance of the source-domain validation performance and target-domain test performance. During the research, there is no training or tuning on target-domain data, and all the experimental scripts, split of datasets, and assignment of classes are recorded to ensure reproducibility.

Together, these preliminaries establish the conceptual and experimental foundation for evaluating zero-shot device generalization using meta-learning in skin lesion classification.

2.2 Review of Existing Research

Automated skin lesion classification has evolved rapidly with advances in computer vision and deep learning. The initial methods depended upon handcrafted features and old fashioned classifiers, and they centered on color histograms, texture descriptors, and border anomalies. As large-scale dermoscopic datasets became more available and the training based on the use of a GPU became sufficiently mature, the direction of research moved to deep convolutional neural networks (CNNs), which proved much more effective in the conditions of controlled settings [4][1]. Although they perform well on within-domain tasks, a weakness that has been observed to reoccur with these systems is their poor generalization with imaging domains. Models that are trained on dermoscopic images do not work often when tested on images obtained with other devices, especially smartphones [14]. This degradation is mainly due to domain shift due to variation in color rendition, resolution, illumination, compression artifacts and acquisition protocols. Consequently, it causes models to learn device-specific patterns rather than intrinsic lesion characteristics, reducing the applicability of models in real-world deployment scenarios [9][11]. In order to solve these challenges, recent studies have explored few-shot and meta-learning methods in the analysis of dermatological images. Meta-learning models seek to enhance flexibility by learning models on distributions of tasks instead of on fixed datasets. It has been shown in several studies that episodic training has the ability to enhance robustness in low-data and cross-domain applications, such as those to skin disease classification [13][3]. Prototypical Networks have been demonstrated to deliver stable and interpretable performance when used in the medical imaging environment as a result of the metric decision structure [3]. More expressive types of meta-learning like Meta-Baseline [6], FEAT [17] and MetaOptNet [8] have also extended this paradigm to learned classifiers, embedding relying on attention or maximization of objective functions. Parallel Domain shift and domain generalization studies are well-represented in medical image analysis. Several of the available methods emphasize domain adaptation, which presumes the availability of target-domain information during training in either labeled or unlabeled form [11]. These methods proved useful in particular settings, but not so much in clinical environments, in which the lack of available target-domain data is a problem of privacy, cost, or regulation. Experiments on cross-domain few-shot learning have pointed to the fact that the performance on the zero-shot task is significantly worse when compared to within-domain performance, further demonstrating the challenge of generalization in the absence of target exposure [9]. A different strand of research has explored the smartphone-based and mobile dermatology applications with the encouragement of the rising availability of consumer-grade imaging devices [14]. The studies highlight the difference between smartphone and dermoscopic images and most of them have revealed extreme decline in performance when models trained on dermoscopic image datasets are transferred to mobile images. Such works show that it is possible to analyze skin lesions with mobile devices, however, most of them use fine-tuning or implicit adaptation to the target domain, which prevents their application in the context of strict zero-shot. In a bid to reduce variability due to devices, various researches have investigated data augmentation and device simulating methods, such as color perturbation, compression artifact blurring, and style transfer. These methods are meant to show a broader variety of visual appearances to the models during training which enhances robustness [32]. Even though augmentation strategies can help mitigate sensitivity to acquisition conditions to some degree, they do not help mitigate the underlying difficulty of acquiring representations that generalize to unseen devices. Last but not the least, the significance of evaluation protocols, reproducibility, and interpretability in medical imaging research has also been highlighted in the recent literature. Splits of datasets in a manner that is inconsistent, implicit exposure to target domains, and use of accuracy as a single measure has been defined as major sources of overly optimistic performance reporting. A number of surveys and benchmark studies recommend rigid source-target domain separation, the application of metrics based on class balancing like macro-averaged F1-score, and disclosure of the experimental environment [16][20][28]. All the literature together emphasizes the progress achieved in the automated classification of skin lesions but also reveals a critical gap. In spite of the fact that meta-learning, domain generalization and augmentation-based strategies had been investigated separately, there are no empirical studies of a strict-environment evaluation of meta-learning in a zero-shot and cross-device setting using smartphone images as an entirely unknown target domain. This thesis fills this gap by exploring systematically the hypothesis: episodic meta-learning can help to prevent domain-induced failures without target-domain exposure of any kind.

Few-Shot and Meta-Learning in Dermatology

Meta-learning in medical imaging is also a significant move in the direction of overcoming the lack of annotated data and the problem of class imbalance that constitute two key issues in dermatology. Unlike in the context of traditional supervised training where the objective is to induce high performance on a small and pre-set group of classes, meta-learning is structured in such a way that it learns models that are capable of making predictions on novel classes or sub-domains with a small number of labeled data. The paradigm is often adopted as episodic training, which is also very common in real-world clinical settings where few-shot learning conditions frequently arise [3][2]. A pioneering and well-known work in the field is Meta-DermDiagnosis by Mahajan et al. [12], which showed that episodic meta-learning can have a beneficial impact on the classification of infrequent dermatological diagnoses. The model trained on simulated few-shot tasks acquired more generalizable visual representations than traditional convolutional neural networks, especially in the case where class-level data were small. This article has formed a background in further studies investigating meta-learning-based methods in identifying skin diseases and was often cited as among the earliest effective uses of meta-learning in dermatology [13][28]. Building on this idea, the Few-Shot Classification with Multiscale Feature Fusion framework [32] added multi-scale feature representations to the episodic pipeline, including local texture and global lesion shape information. Such a strategy improved the morphological feature representation that is highly essential in the differentiation of lesion classes that are visually close together like that of melanoma and benign nevi. The authors found that better recall was observed in poorly represented lesion categories, which demonstrates the usefulness of hierarchical feature extraction in the few-shot dermatologic classification problem. The original Prototypical Network formulation was further adapted specifically for skin lesion classification [30]. This method was using an embedding function that was trained to reduce the distance between same-class and reduce the distance between different classes. The mean of the support embeddings were then computed as the class prototypes and the query samples were classified by distance based decision boundaries which were based on either Euclidean similarity or cosine similarity. This formulation, which is based on metrics, has high interpretability since prototypes can be visualized or directly compared, and has a stable training behavior. Prototypical Networks are particularly suitable to dermatological applications because these properties are important especially in dermatology where transparency and robustness are vital [3]. Further extensions of the Prototypical Network framework were suggested in the form of Influential Prototypical Networks [18] which added prototyping weighting schemes to help counteract the influence of noisy and heterogeneous data. This method increased resistance to outliers and made the method more resilient to annotation noise by downweighting atypical or mislabeled samples. These properties are particularly applicable in dermatology where intra-class variation may be large because of illumination, acquisition, or disease progression stage variations. Metalearning has been further applied to multi-task diagnostic and mobile diagnostic contexts in addition to single-task classification. Multi-Task and Few-Shot Learning-Based Mobile Diagnosis Platform [25] combined meta-learning and multi-task learning and trained one model on several similar diagnostic tasks. The approach led to the creation of more transferable representations which were generalized to new tasks and showed the practicality of meta-learning as used in smartphone-based teledermatology systems. These results can be interpreted as indication that meta learning could have a role to play in facilitating class adaptation as well as real-world implementation conditions with mobile devices. Overall, the analyzed literature shows that meta-learning is effective in increasing data efficiency and improving the performance of minority and underrepresented classes. But there is a significant limitation: most of the current research analyzes the problem of meta-learning in comparatively weak conditions, typically testing its performance on the same dataset (e.g., HAM10000) or with access to target-domain data upon adaptation. Consequently, the performance of meta-learning systems in the presence of severe domain change, between imaging devices, is not well studied [9] [11]. The thesis is based on the meta-learning paradigm, and it specifically fills this gap. Episodic tasks are designed in a device-conditioned fashion, instead of just learning a few-shot class adaptation, with support and query samples based on dissimilar dermoscopic source dataset. This architecture forces the network to acquire device-free representations and it mimics the existence of hidden acquisition devices. This way, the proposed strategy builds on the theoretical advantages of Prototypical Networks and targets the problem of generalizing to real-world smartphone images that is highly untested, directly.

Domain Shift and Domain Generalization

Domain shift phenomenon is a phenomenon whereby the distribution of data in test-time is not the same as in training, making the performance seriously degrade. These changes are common in dermatology when comparing datasets of the same dermatologist with different dermatoscopes, different clinical centers or when the acquisition conditions are different. Systematic variations in imaging devices bring about these changes in color calibration, illumination, magnification, compression, and framing, and even visually identical lesions differ statistically across datasets. Thus, the reliability of generalization of models trained in controlled dermoscopic situations to images taken in various conditions tends to be low. In order to deal with this problem, both domain adaptation (DA) and domain generalization (DG) methods have been studied. Domain adaptation presumes that data in the target domain - usually unlabeled - is available to the training procedure, whereas domain generalization aims to achieve robustness regardless of target-domain exposure. This fact is especially important to this thesis that takes a stern zero-shot paradigm, at no time are samples of the target-domain provided. One of the most detailed analyses of domain adaptation methods used to dermoscopic images is the benchmark study Mitigating the Influence of Domain Shift in Skin Lesion Classification [31]. The authors explore the concepts of adversarial alignment, CORAL-based feature matching, as well as style-transfer methods in relation to different sources in the ISIC. Although these techniques are effective in cases where the target domain samples are not labeled, they become ineffective with an increasing gap in domains between the two domains or when the samples of the target are not known. These results reveal a primary weakness of adaptation-based techniques in the face of literally invisible domains, which is typical in practice. In order to address these shortcomings, domain generalization frameworks have been suggested. Khandelwal and Yushkevich presented Domain Generalizer [10], a few-shot meta-learning model designed to be used in medical imaging DG. Their algorithm picks source domains episodically, meta-trains to some and meta-tests to others in order to promote invariance to domain-sensitivity to features. This approach is related to the methodological path of this thesis, which cites device variation as a latent variable to be generalized over. Other work in the domain of meta-optimization, such as Meta-learning with Implicit Gradients [22] and Few-Shot Learning with Medical Imaging: A Comparative Analysis [26], also indicate that inner-outer loop optimization can facilitate environmentally and distributional-invariant representations. However, regardless of such developments, the domain generalization studies on dermatology are limited. The majority of the studies are based on dermoscopic data sets that were obtained in reasonably similar clinical practice and fail to explicitly provide consideration of the increased variability provided by consumer-grade smartphone imaging [24][34][3][11]. Additionally, the mixed datasets are often reported in evaluation protocols with average accuracy, which may hide class-specific and device-specific degradation of the performance [31][5][1]. One of the crucial constraints of the current DG and meta-learning methods is that the variation of devices is not often explicitly imposed on the level of tasks-construction. Whereas the previous approaches promote the extrapolation across datasets, they do not normally build device-sensitive episodic tasks, where support and query examples are purposely sampled across dissimilar acquisition devices. Consequently, models can still be dependent on device-related clues instead of acquiring a representation that is actually device-invariant [13][6]. This gap is directly addressed by this thesis, which considers a device-aware episodic training strategy, where the support and query sets are sampled across the various source datasets in each episode. This episodic construction is explicitly used to simulate the difficulty of unseen acquisition conditions by making the model do classification across device boundaries during training. This architecture builds on earlier domain generalization meta-frameworks [10][26][28] and is consistent with recent evidence that task-based variability is a compelling force in learning of robust representations [13][5][2]. Further, unlike adaptation-based models, the suggested framework is trained on HAM10000 and BCN20000 dermoscopic datasets and tested on the PAD-UFES-20 smartphone dataset that will not be seen at all during training, validation, and hyperparameter optimization. This design isolates the actual effect of device shift and prevents the confounding effect of exposure to target domain. In short, although the current state of DA and DG approaches has improved the knowledge on distributional alignment in dermatological imaging, they cannot meet the requirements of real-life cross-device deployment. This thesis offers a more realistic and rigorous picture of evaluating robustness in the case of extreme dermoscopy-to-smartphone domain shift by adjoining domain generalization principles with explicit device-aware episodic construction and zero-shot evaluation [13][5][3][19][27][4][1].

Smartphone Dermoscopy: Image Quality and Artifacts

The increasing popularity of smart phones has greatly extended the process of teledermatology where both patients and clinicians are able to take pictures of skin lesions without having to use special imaging devices. Although this availability has enhanced accessibility and convenience, it has also brought a lot of variability in image quality as compared to the traditional dermoscopic imaging. Switching to uncontrolled consumer smartphone photography leads to greater heterogeneity due to variations in camera sensors, white-balance algorithms, compression settings, focal length and user practices of acquisition. All this is what leads to the device gap which is the foundation of the present study. It is consistently proved by empirical evidence that deep learning models that are trained on clinical dermoscopic data do not generalize to smartphone-acquired images. A paper that assessed a neural network implementation on smartphones [34] demonstrated that the diagnostic accuracy of the same algorithm significantly decreased when it was used on consumer devices. The experimentally determined decrease in performance was explained by environmental lighting contrast, motion-blur, and imbalanced framing, which demonstrates that high laboratory performance would not always be correlated with trustworthy behavior in the real world. To overcome these shortcomings, the use of dermoscope attachments in smart phones has been studied in a number of studies. The study by MobileSkin research [24] was conducted to explore the efficacy of dermoscopes on mobile phones to improve the quality of the image. Although this kind of attachment enhanced resolution and control of brightness, noises and artifacts were still left behind. Consequently, the accuracy of the models was more enhanced compared to raw smartphone photography but was still not as high as the one of standard dermoscopic images. In the same manner, The Role of an Inexpensive Smartphone Device in Teledermoscopy [21] has shown that low-cost attachments could be used to provide images of good enough quality to perform simple lesion triage, but still there were artifacts in the color that still disrupted convolutional neural network predictions. Systematic studies on image quality optimization have also been used to give additional insights. In Optimization of Digital Image Quality with Better Skin Cancer Detection [33], the authors tested the effect of varying the image resolution, compression ratios and calibration of color on the performance of deep learning by experimentally adjusting these parameters. They found that aggressive JPEG compression and low-contrast imaging has a strong detrimental effect on the sensitivity of the classifier especially when detecting melanoma, where the fine-grained color graduates are a critical diagnostic feature. The findings highlight the importance of device-level artifacts in feature extraction and also should be directly modeled or countered during training. The use of auxiliary information to offset the visual heterogeneity has also been studied recently. In A Deep Learning-Based Multimodal Fusion Model of Smartphone-Collected Clinical Images [23], multimodal fusion models were applied to integrate image data with the metadata of patient demographics and lesion location. Although this method made the models much more robust compared to those based on image only, it was in essence limited by the quality of the input visuals. Metadata could not compensate for low image fidelity completely, which supports the significance of dealing with the artifact of devices at the level of representation. All these studies lead to the same conclusion that smartphone imaging generates a significant and systematic domain shift that the existing supervised models cannot manage effectively. The variability of devices is not some nuisance but an essential obstacle to making automated systems of skin lesion classification usable in any real-world setting. This literature base directly supports the fact that PAD-UFES-20 can be effectively used as a serious zero-shot assessment benchmark [1] and inspires the creation of device-sensitive learning models that promote resistance to invisible phone features. This approach to the domain shift will specifically seek to fill the gaps in current methods of supervised approaches.

Data Augmentation and Simulation for Robustness

One of the most widely used methods of enhancing cross-domain robustness is data augmentation, which trains models on a larger variety of data. The purpose of augmentation techniques is to widen the effective training distribution by modeling transformations which might happen under unknown conditions of acquisition. Augmentation has two complementary roles in dermatological image analysis: augmenting the diversity of datasets by overfitting, and directly modeling variations in images caused by devices that are different in imaging hardware and settings. The main geometric invariance method of classical augmentation like rotation, flipping, cropping, and scaling has been popular in the early CNN-based skin lesion classifiers. These changes however, have little effect in the minimization of appearance discrepancies associated with the differences between imaging devices since they do not significantly change color allocation, light, or texture statistics. Consequently, later literature has focused on photometric and appearance-based augmentation methods which are more realistic to the system variability in the real world. The authors of the study Optimizing Digital Image Quality for Improved Skin Cancer Detection [33] have been able to systematically investigate the effects of the resolution, JPEG compression, contrast, saturation, and color calibration on the performance of deep learning. Their experiments showed that differences in compressions and color balance had a major influence on the classifier sensitivity especially in melanoma, which is associated with small color gradients that are critical in diagnosis. Such results suggest that perturbations of photometrics can be used to simulate differences between capture devices in a meaningful manner and should be involved in training. Benchmark works on domain shift in skin lesion classification [31] also found similar results and tested several domain adaptation and augmentation strategies on ISIC datasets. That paper demonstrated that color jitter, histogram equalization and gamma correction yielded moderate cross-dataset transfer enhancement, but excessive aggressive transformations caused worse lesion visibility and worse performance. The authors of the article Optimizing Digital Image Quality for Improved Skin Cancer Detection [33] performed the systematical analysis of the effect of resolution, JPEG compression, contrast, saturation and color calibration on the performance of Deep learning. Their experiments revealed that the disparity in the compression level and color balance played an enormous role in the sensitivity of the classifier particularly with melanoma wherein the color gradients are diagnostically meaningful. The results of such nature lead to the fact that photometric perturbation is significant in simulating discrepancy between capture devices and it should be taken into consideration in training. Similar conclusions were drawn in benchmark studies of domain shift in skin lesion classification [31], by evaluating different domain adaptation and augmentation algorithms on ISIC datasets. An article showing that color jitter, histogram equalization, and gamma correction seemed to cause moderate increases in cross-dataset transfer, but grossly aggressive changes worsened the look of lesions and impaired the results was published that year. Following on these results, other studies have discussed the use of device-oriented augmentation as domain generalization, as opposed to domain adaptation. More recent work like Robust Domain Generalization via Style Randomization [13] and Domain Generalization with MixStyle [5] has shown that no-target domain random perturbation of style statistics can make images more resilient to unseen domains. Though these techniques have not been created with the explicit purpose of dermatology in mind, their findings are consistent with those of the medical imaging that global appearance statistics are highly effective in shaping model behavior. The further extension of this concept is style transfer-based augmentation methods that transform the appearance of images without altering the structure of lesions. Modes of separation of content (lesion morphology) and style (color, texture, illumination) are achievable using techniques based on Adaptive Instance Normalization (AdaIN) and style manipulation using GAN. Although most style transfer methods use reference images in the target domain, randomized or source-only style sampling has been suggested as a method of making predictions about unknown device properties. The results of other studies, including MixStyle [7] and Cross-Domain Style Augmentation for Medical Images [1], demonstrated that feature-wise mixing of statistics in training can yield a strong domain robustness without explicit target-supervision. The significance of modelling device artifacts in dermatological use is also further highlighted in smartphone imaging research. Studies on smartphone-based diagnosis systems [24][34][21] regularly indicate that compression artifacts, haphazard lighting, and color variation depending on sensors are significant causes of performance deterioration in deploying models trained on dermoscopic data. These results guide the incorporation of device-conscious augmentation techniques, which comprise color perturbation, JPEG compression, and blur, that are simply adopted in this thesis as techniques of simulating smartphone-specific artifacts during training. In addition to augmentation, large-scale pretraining and representation learning frameworks give more robust information. Multi-Aspect Knowledge-Enhanced Vision-Language Pretraining framework [35] proved the fact that when various visual and semantic distributions are exposed, the strong capability of generalization on a zero-shot may be obtained. Equally, self-supervised and foundation-model methods developed more recently [3][6][17] indicate that representations trained using large and diverse datasets are less prone to domain-specific noise. Although these methods are computationally demanding and may not be practical when conducting research at the undergraduate level, they support the main hypothesis that diversity of appearance during training is an essential factor in cross-domain generalization. Augmentation in this thesis is not considered individually but in combination with episodic meta-learning to make it all the more robust. The intra-episode variability is augmented, and the episodic training imposes the generalization across the tasks built by using other source datasets. The motivation behind this combination is related to the previous meta-learning in dermatology [12][30][18] and the domain generalization frameworks [10][26], except that the technology targets specifically device-induced appearance shifts. The combined application of device-targeted augmentations and meta-learning episodes is a poorly studied area of existing literature and serves as a methodological contribution of this research.

Evaluation protocols, reproducibility, and interpretability

Strict testing procedures are the basis of plausible assertions on the generalization performance in medical image analysis. Bad splits of datasets, inappropriate choice of metrics, or leakage between training and testing domains may result in exaggerated performance estimates and false inferences. Previous research in dermatology has consistently focused on the need to have reproducibility and strict experimental separation in order to have valid, cross-study comparisons. Both the Benchmark Study on Domain Shift in Skin Lesion Classification [31] and Meta-DermDiagnosis [12] clearly embrace reproducible evaluation procedures, fixed random seeds, publicly available split definitions, and distinct separation between training, validation, and test domains. Such practices permit open benchmarking and minimize vagueness in presented outcomes. More recent reproducibility-intensive studies in medical AI further support the claim that the change in performance due to various data splits or evaluation procedures can be just as large as the improvements due to algorithmic innovation [13][5]. The suggested observation highlights the importance of regular evaluation pipelines, especially when the domain generalization is the goal, and the performance is extremely sensitive to the composition of the dataset. Thus, this thesis adheres to a well-defined and reproducible protocol: datasets are split in a predetermined way, and the same splits are applied to all models, and no target-domain images are involved in any step of training, validation, hyperparameter optimization, or model selection. Another important element of sound assessment is evaluation metrics. Even accuracy alone is not generally thought to be sufficient to the medical classification problem because there is a devastating imbalance of classes and even clinical significance of the various classes is uneven. Previously, it was always advised that macro-averaged F1-score, per-class recall and ROC-AUC are more informative indicators of clinical utility [12][31][26][7]. Macro-F1 does not give more weight to the dominant classes, so that the high-performance of the dominant classes like benign nevi does not conceal the poor performance of the more uncommon but clinically important classes like melanoma or squamous cell carcinoma. These reasons are directly translated into the assessment plan that is chosen in this thesis because macro-F1 is the main measure and the accuracy is presented only as an auxiliary measure. Beyond quantitative metrics, interpretability has become increasingly important for clinical trust and deployment. It is critical to understand the reasons why a model leads to a certain prediction to identify failure modes especially when there is a domain shift. The use of visualization methods like Gradient-weighted Class Activation Mapping (Grad-CAM) has become very common in dermatological deep learning as a means to determine whether models prioritize lesion areas over background artifacts. The SkinLesNet research [29] evidenced that Grad-CAM visualizations can be used to detect whether CNNs are able to understand clinically significant features or arbitrary spurious relationships. On a similar note, in multimodal fusion methods [23], Grad-CAM was used to ensure that the enhancement of performance was brought about by lesion-centric attention and not an effect of contextual bias. Recent interpretability works also indicate that the domain shift may result in attention drift and that models trained on dermoscopic image may fail to attend to the foreground texture or light artifacts when tasked with smartphone image classification [1][3]. The failures are hard to measure by relying on scalar measures. Qualitative interpretability analyses, therefore, qualify as a cross-validation method, which gives confirmation that measured increases in performance are related to meaningful visual reasoning. Selective application of Grad-CAM visualization is used in this thesis to confirm lesion-centered attention patterns and fail cases during cross-device testing. One of the major ideas that have come about as of the recent literature is the concept of robustness gap, which is the decrease in performance between target-domain testing and source-domain validation. Several works note that macro-F1 decreases by 20-50% during the transfer of models between dermoscopic data and smartphone images [24][33][34][6]. By measuring this gap directly, domain sensitivity is measured and different learning paradigms can be compared not by absolute accuracy values. In that regard, strong robustness will be directly quantified in the given work, as the percentage loss difference between the mixed source-domain verification (HAM10000 + BCN20000) and the target-domain performance in PAD-UFES-20. This thesis is consistent with best practices set in previous studies by incorporating strict evaluation procedures, reproducibility methods, clinically significant measures, and interpretability studies and overcomes their shortcomings. Specifically, as opposed to most existing experiments, which average performance between datasets or use partial target exposure, the current paper examines actual zero-shot generalization at a fully unseen device test. The resultant experimental design offers a clear and justifiable foundation of measuring meta-learning as a significant enhancer of cross-device robustness in skin lesion classification.

Synthesis: Gaps and Where This Thesis Fits In

The reviewed literature in this chapter shows that the automated skin lesion classification research is active and varied, but is segmented into various partially overlapping directions. The effectiveness of meta-learning and few-shot learning has been demonstrated to be very high in situations with minimal amounts of labeled data and class imbalance, which are typical dermatology problems. Nevertheless, these techniques are not often tested through extreme cross-domain shifts, especially the ones that include changes of dermoscopic pictures into pictures obtained with the help of a smartphone. The majority of the currently known literature conducts meta-learning in a single dataset or with slightly different conditions of acquisition, and the major issue of true device-level generalization has not been investigated [12][30][32][13]. The domain adaptation and domain generalization studies have made great contributions to the study of distributional alignment and feature invariance. However, in most domain adaptation methods, unlabeled or partially labeled target-domain samples are accessible at training, which is frequently not realistic in the medical deployment case [31][5][7]. Even domain generalization experiments often use several dermoscopic datasets taken in close clinical conditions which make it difficult to represent the variability brought about by consumer-grade smartphone imaging [10][26][1]. Due to this fact, robustness claims are also frequently restricted to controlled experimental environments and not real-world implementation. Research involving smartphone dermatology puts definite records of the degree of variability presented by variations in camera sensors, illumination, compression, and user behavior [24][33][34][21][23][3]. In all these works, there is always a significant report of performance regression on smartphone images when models that had been trained on dermoscopic data are used. Most smartphone-centered studies however focus on enhanced acquisition devices, attachment-based dermoscopy, or quality improvement of images instead of algorithmic based approaches which are run in a zero-shot setting. In turn, these researches fail to offer a principled learning paradigm that will be able to address the unseen device properties. Mechanisms that have been put forward to augment robustness are data augmentation and style-transfer techniques . They expand the variety of training distributions [31][33][35][6]. Although these methods enhance tolerance to photometric and textural changes, they are generally used as a part of standard supervised learning pipelines and are not systematically combined with task-level frameworks of learning like meta-learning. In addition, most of the style-transfer methods use reference images on the target side thus making them less applicable to actual zero-shot situations [17][8]. Lastly, reproducibility, interpretability, and robustness measures are increasingly becoming important in current AI medical literature, but their use is not consistent across the studies [12][26][31][5]. Specifically, the concept of robustness is frequently implicitly reported based on aggregate accuracy, and not explicitly shown as a source to target performance drop. Interpretability analyses are often not mandatory or quantitative and therefore their contribution to justifying domain-invariant reasoning is limited. This thesis unites these earlier strands that were severed into a single and flawless evaluation structure. It uses Prototypical Networks [30][18] as a stable, interpretable and well established meta-learning, which has shown good results in dermatological few-shot classification [12][32]. The proposed approach is based on the domain generalization meta-frameworks [10][26], where each episode is based on support and query samples that are sampled across various source datasets. This design specially promotes the acquisition of representations that are independent of device-specific acquisition properties. Moreover, application specific extensions to the methodology are made based on empirical observations in the literature of smartphone dermatology [24][33][21][23][34][3]. The latter synthetically recreates unseen device conditions without using target-domain samples, which is in line with the recent results that controlled exposure to diversity enhances robustness [35][6][2]. In contrast to the large-scale methods of pretraining which involve significant computational costs [35][15], this thesis implements a lightweight but conceptually sound methodology that integrates episodic design with device-specific variability. The experimental analysis is based on high benchmark guidelines [26][31][5], where PAD-UFES-20 is never seen in training, validation, or model selection. Generalization is evaluated holistically by macro-F1, and explicit robustness-drop quantification which quantitatively measures domain shift sensitivity. All in all, the thesis adds a coherent, reproducible and deployable generalization of the skin lesion classification in zero-shots to a device. It combines the meta-learning theory, the domain generalization principles, and the smartphone imaging research experience to provide the answer to the significant and poorly investigated issue: allowing the use of models trained on clinical dermoscopic data to process real-world smartphone images without any retraining or target-domain exposure.

2.3 Summary of Key Findings

In the studies considered, several consistent trends are apparent when it comes to the efficiency and the constraints of the current strategies of automated classification of skin lesions under the conditions of scarce data and varied domains.

- Meta-learning methods, especially those based on Prototypical Networks, and gradient-based models such as MAML, are consistently described to exhibit better transferability and enhance recall on under-represented lesions classes than traditional single-task supervised training [12][30][32][18][13][5]. Models can learn task-adaptive representations with greater cross-class distribution generalization by using episodic training that models task-level learning. Those properties have projected Prototypical Networks as computationally effective, steady, and explainable bases of dermatological few-shot learning [30][18][7].

- Domain shift remains a dominant factor limiting cross-device performance. The models trained on dermoscopic data always have extensive degradation on smartphone images because there are systematic differences in sensor response, illumination, compression, framing, and user behavior [24][33][34][21][23][3]. Although unsupervised domain adaptation (UDA) and domain generalization (DG) methods partially decrease this gap, they rarely eliminate zero-shot differences, especially when the data of the target domain is not replied to [31][5][1][6]. The implications of these findings are that it is highly important to employ stringent cross-domain evaluation procedures that entail isolating genuine generalization capacity.

- All studies of smartphone and field dermoscopy have empirically shown poorer diagnostic accuracy than controlled dermoscopic settings. This degradation is majorly associated with heterogeneity of sensors, uneven acquisition pipelines [24][34][3]. These facts correspond to the necessity to use real-world images like PAD-UFES-20 and encourage the necessity to use synthetic device-variation modeling to estimate the noise and color distortions on a field scale during training [33][17].

- It has been demonstrated that data augmentation and style-transfer are useful methods of increasing the training distribution, and enhancing stability against appearance variation [31][33][35][6]. Carefully parameterized perturbations e.g. color shifts, blur, compression, etc, enhance cross-domain robustness, and uncontrolled augmentation can corrupt lesions semantics and damage clinical validity [33][8]. Multimodal and vision-language pretrainers Large scale Multimodal and vision-language pretrainers also suggest potential possibilities of zero-shot transfer, but are computationally intensive and not easily executable in limited research environments [35][15][16].

- The methodology of evaluation is very important in the validation of claims on generalization. Measures of macro-averaged F1-score, class-specific recall and explicit measures of robustness-drop are more diagnostic than overall score, especially when there is some form of class imbalance [12][26][31][5]. Reproducibility practices including fixed target splits, shared label mappings and consistent random seeds are necessary but used inconsistently in the literature [26][31][1]. Interpretability studies on methods like Grad-CAM also indicate whether models focus on clinically significant parts of the lesions or rely on device-specific artifacts, thus revealing shortcut learning when faced with change of domains [23][29][16][20].

- There are three complementary generalization methods that are repeated across the literature: meta-training to promote such representations, augmentation to increase domain coverage and adaptation mechanisms based on unlabeled target data [10][26][31][5][6]. Although metadata fusion has been demonstrated to increase robustness in certain smartphone-based systems, it is not tested in the case of hard zero-shot constraints and is incapable of completely addressing a degenerate visual input [23][3][9].

- A number of research gaps are still present. The literature has not yet demonstrated meta-learning based on very strict zero-shot device requirements, there are not enough reproducible cross-device benchmark procedures, interpretability analysis is not consistently used, and the explicitly device-aware episodic construction is not explicitly used in dermatological meta-learning. These restrictions make many of the reported robustness claims practical.

All these together have inspired the current thesis. This paper provides a combined study of a principled meta-learning baseline Prototypical Networks [30][18] combined with device-aware episodic sampling and application-specific augmentation approaches, and tested in a strict zero-shot PAD-UFES-20 protocol. The experimental design determines the effect of episodic simulation of cross-device variance in reducing robustness loss, and maintaining clinically significant patterns of attention. They use reproducible splits, evaluation scripts and consistent metrics to enable transparency and allow future benchmarking, which helps overcome some of the key limitations found in the existing literature.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The main goal of this study is to assess whether meta-learning approaches have the capability of enhancing zero-shot generalization in the process of skin lesion classification in the case of severe cross-domain shift. The system is meant to be used in a zero-shot strict scenario whereby the models are only trained on dermoscopic images and are only tested on smartphone-captured clinical images but not on the target domain images at all during the training, validation, and model selection. Every model follows one set of six-class diagnostic mapping and a standard preprocessing pipeline in order to be able to make fair and interpretable comparisons. The design of the experiment needs that the feature representation (backbone architecture) and the learning paradigm (supervised or meta-learning) are separated to allow the controlled study of their independent and interactive effects on robustness. Macro-averaged F1-score is used to evaluate performance because it treats disparities in classes as a primary issue, and because it underlines the strength of results in every diagnostic category and not on the highest accuracy.

3.2 Societal Impact

This study fills a gap of critical difference between benchmark medical image classification performance and operating conditions in real-life deployment. The study is realistic in clinical and consumer use scenarios by generalizing dermoscopy to smartphones since no constraints on image capture are imposed, and no expert annotation exists. Improvements in zero-shot robustness could support more reliable decision-support tools in resource-limited or remote settings. The proposed system is to be used as a decision-support device and not a substitute of professional medical judgment. Computer forecasts are placed in a supplementary role instead of being in official opinion-giving, which validates the need to have clinical supervision in any medical decision-making.

3.3 Environmental Impact

The main issue of environmental impact of this study is the use of computational resources to train and evaluate the model. The use of efficient backbone architectures to allow broad comparison of experiments and prevent hyperparameter sweeps or repeated large scale execution is preferred by the experimental design to minimize unnecessary energy consumption. The study aims to bring an equilibrium between the scientific rigor and responsible computational practice by laying greater emphasis on controlled experimentation in lieu of brute-force optimization.

3.4 Ethical Considerations

All of the experiments are carried out on the data, which are publicly available and de-identified, and comply with the principles of data privacy and ethical research. Automated medical decision systems are exposed to ethical risks, which are effectively overcome by focusing on the characterization of failures and robustness analysis, instead of deployment claims. The paper defines that all the proposed models should always be considered as supporting tools of decision-making and does not substitute clinical judgment, especially in extreme circumstances of domain shift when uncertainty in prediction is high. This work does not present deployment-ready diagnostic systems but rather analyzes constraints and malfunction patterns and assesses the meta-learning practices as the possible way to enhance robustness and be able to apply the associates safer in the future.

3.5 Project Management and Risk Considerations

This thesis is divided into serial and non-concurrent phases to provide clarity of the methodology and effective time and calculative resources management. The main stages of management involve:

- Performing an extensive literature review of meta-learning and cross-domain medical image analysis.
- Training and preprocessing of dermoscopic and smartphone image datasets based on a single diagnostic mapping.
- Implementing selected backbone architectures and meta-learning systems.
- Executing controlled experiments under strict zero-shot setting
- Comparison of results and analysis of meta-learning methods and conventional supervised training.
- Documenting findings and preparing the thesis report.

With time and calculation resources, experiments are ranked in terms of anticipated methodological worth. The risks associated with underperformance of models, resource constraints, and the instability of the experiments are addressed with the help of controlled comparisons, regular evaluation procedures, and close documentation.

3.6 Economic Constraints

The project has a low economic footprint as it uses publicly accessible datasets and open-source applications. The first expense is computational cost, and it is controlled using representative model configurations and redundant experiments are avoided. This cost-effective measure will make the study viable within the funds at their disposal and yet yield informative information that can be used in actual research and implementation processes.

Chapter 4

Proposed Methodology

4.1 Methodology Overview

This chapter introduces the research methodology to be followed in this study towards the classification of zero-shot skin lesions under severe cross-domain shift. The suggested methodology analyzes the hypothesis of whether meta-learning techniques can enhance robustness in the cases when models trained on dermoscopic images are applied to clinical images collected on smartphones and without access to target-domain images.

Majority of the prevailing medical image classification researches assume that training and test data is gathered under comparable conditions of acquisition. Conversely, this is aimed at a more realistic deployment setting, whereby dermatological decision-support systems conditioned on high-quality dermoscopic images are used to apply to consumer-quality smartphone images taken in unconstrained conditions. This distortion brings in a great deal of variation in the illumination, resolution, background content and color distribution that can be greatly detrimental in the performance of conventional supervised models.

To overcome this issue, the implementation of the suggested design combines meta-learning, multiple backbone architectures, and a rigid zero-shot assessment protocol into a single experimental procedure. This is aimed not just at evaluating classification performance, but also at determining the effect of various learning paradigms and representation decisions on generalization in the case of severe domain shift.

4.1.1 Motivation, Problem Context, and Design Principles

Skin cancer is also one of the most common cancers in the world, early diagnosis is a very important factor in enhancing patients and related outcomes. The recent developments in deep learning have also made image classification models perform well in dermoscopic images datasets under controlled experimental settings in certain instances surpassing even expert dermatologists. Nonetheless, this kind of performance is usually based on training and testing data that have similar characteristics of acquisition.

Real-life clinical practice has seen photographs taken by untrained people on smartphones and under different lighting conditions without standardization of imaging practice. These images are quite different compared to dermoscopic images in terms of texture, color fidelity, scale and background noise. Consequently, dermoscopy-based models that are trained using dermoscopic datasets often provide inaccurate predictions when presented with images obtained using a smartphone.

This gap between benchmark performance and deployment reality motivates the central research question of this thesis: Can meta-learning improve zero-shot generalization for skin lesion classification under severe dermoscopy-to-smartphone domain shift? Instead of solving this issue by means of domain adaptation or fine-tuning to the target domain, this paper uses a strict zero-shot context where no target-domain samples are provided throughout the training. Such formulation is indicative of the situations in which the labeling of target-domain data is not feasible, expensive, or ethically prohibited.

Three principles are used to design the proposed approach. One, it maintains a very rigid separation between source and target domains. Training and validation processes are provided with the use of only dermoscopic images, and all the assessment is carried out with smartphone images that are not seen during training. It does not allow any fine-tuning, calibration, or episodic exposure to the target domain, unlike those studies that make weaker generalization assumptions.

Second, the approach takes an explicit factorization of feature representation and learning paradigm. The models are all broken down into a backbone architecture that extracts features and a learning strategy, either a traditional supervised learning strategy, or a meta-learning strategy. Such isolation allows the controlled comparisons to isolate the contribution of performance differences by representation quality, learning paradigm or the interaction of both.

Third, the methodology focuses on the breadth of methodology and depth that is controlled. A wide range of representative backbone architectures is tested to meet the variability of architectures, and specific a priori methods are investigated to analyze various learning paradigms with the help of meta-learning. Further research is only done in cases where it will add value as a balance between experimental and computational capability.

4.1.2 Overview of the Proposed Learning Framework and Scope

The learning model has three phases, namely, source-domain training, meta-learning, and zero-shot evaluation. These phases are indicative of the life cycle of a real deployment scenario.

During the first step, publicly available datasets with dermoscopic images are trained on models. In the case of normal supervised baselines, the optimization process is carried out on mini-batches with the use of cross-entropy loss. In the case of meta-learning methods, episodic sampling is applied in training, in which the model is optimized with respect to a distribution of classification tasks instead of samples.

At the second stage, episodic tasks based on the source domain are used to train meta-learning models. The episodes comprise a support set and a query set, which resemble a classification task. This task-based training introduces the model to format variability and promotes learning representations and inference processes that can generalize to training distributions outside of the training data.

The trained models are tested directly on the smartphone-obtained images of the target domain in a stringent zero-shot protocol in the last step. No fine tuning or adaptation is carried out at this point. The learned representations and inference mechanisms are used to generate predictions and evaluate them using pre-defined evaluation metrics. This phase is the desired deployment situation.

The role of meta-learning in this paradigm is central to it but perceived as a hypothesis as opposed to an assumed enhancement. The methodology is made specifically to evaluate the hypothesis of whether meta-learning can enhance robustness in the presence of extreme cross-domain shift, does not degrade as much as traditional supervised learning, and does not interact with different backbone architectures.

The suggested methodology is prone to practical limitations.Computational constraints limit the set of architectures and learning strategies that can be tried in detail and the strict zero-shot condition prohibits the application of domain adaptation methods that need access to the target domain. These shortcomings are handled by designing experiments carefully with strong emphasis on interpretability, fairness and relevance to real-world implementation, rather than by exhaustively experimenting.

4.2 Dataset Description and Problem Definition

This section gives the description of datasets to be processed in this study and the formulation of the learning problem that is going to be solved by the proposed methodology. As the goal of the given research is to test zero-shot generalization in the situation of severe domain shift, the attention of the experimental design is paid to the correct choice of the data set and strict division between training and testing domains. The specific focus is then placed on the nature of the source and target datasets, their variation in terms of the conditions of acquisition, and the consequences of the variation on the generalization of the models.

4.2.1 Source Domain Datasets

The source domain in this study consists of dermoscopic images obtained from two publicly available datasets: HAM10000 and BCN20000. These datasets are widely used in the literature on automated skin lesion classification and provide high-quality, expert-annotated images acquired using dermatoscopes in controlled clinical settings.

The HAM10000 (Human Against Machine with 10000 training images) dataset contains dermoscopic images collected from multiple clinical centers. Images are captured under standardized conditions using dermatoscopes, which provide magnified and well-illuminated views of skin lesions. Each image is annotated by dermatology experts and assigned a diagnostic label. HAM10000 is commonly used as a benchmark dataset due to its size, annotation quality, and clinical relevance. However, like many dermoscopic datasets, it exhibits class imbalance, with benign lesions such as nevi appearing far more frequently than malignant categories. Despite this imbalance, the dataset remains valuable for training feature extractors capable of capturing fine-grained lesion characteristics such as color variation, border irregularity, and texture.

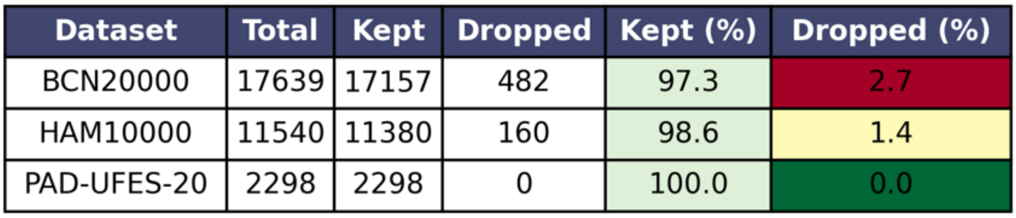
The BCN20000 dataset is another large-scale dermoscopic dataset collected from clinical practice. Similar to HAM10000, images in BCN20000 are acquired using dermatoscopes and annotated by medical professionals. The dataset provides additional diversity in lesion appearance, imaging devices, and patient demographics, complementing the HAM10000 dataset and increasing variability within the source domain.

Rather than training models on a single dermoscopic dataset, this study combines HAM10000 and BCN20000 to form a unified source domain. The motivation for this decision is twofold. First, combining datasets increases the number of available training samples, which is beneficial for training deep feature extractors. Second, it introduces greater variability within the source domain, which may encourage models to learn more generalizable representations. Importantly, both datasets share similar acquisition characteristics, as images are captured using dermatoscopes under controlled conditions. As a result, combining these datasets does not blur the distinction between source and target domains, which remains clearly defined by acquisition modality.

4.2.2 Unified Class Mapping

One challenge in combining multiple datasets is inconsistency in class definitions and annotation granularity. To address this issue, all source-domain images are mapped to a unified six-class labeling scheme, ensuring consistency across datasets and alignment with the target domain. The six diagnostic categories used in this study are:

Actinic Keratosis (ACK)
Basal Cell Carcinoma (BCC)
Melanoma (MEL)
Nevus (NEV)
Squamous Cell Carcinoma (SCC)
Seborrheic Keratosis (SEK)



Dataset	Total	Kept	Dropped	Kept (%)	Dropped (%)
BCN20000	17639	17157	482	97.3	2.7
HAM10000	11540	11380	160	98.6	1.4
PAD-UFES-20	2298	2298	0	100.0	0.0

Figure 4.1: Data Splits

Once the unified six-class mapping has been used, all the datasets of the source domain were recalculated with the help of the unified six-class mapping to clear all the ambiguous, missing, and inconsistent diagnostic labels, which would help assure label reliability. By contrast, the target-domain dataset PAD-UFES-20 was left untouched and no filtering of this dataset was performed to ensure a strict zero-shot evaluation environment. Figure 4.1 presents a summary of the images that were retained and discarded by the datasets after the cleaning process.

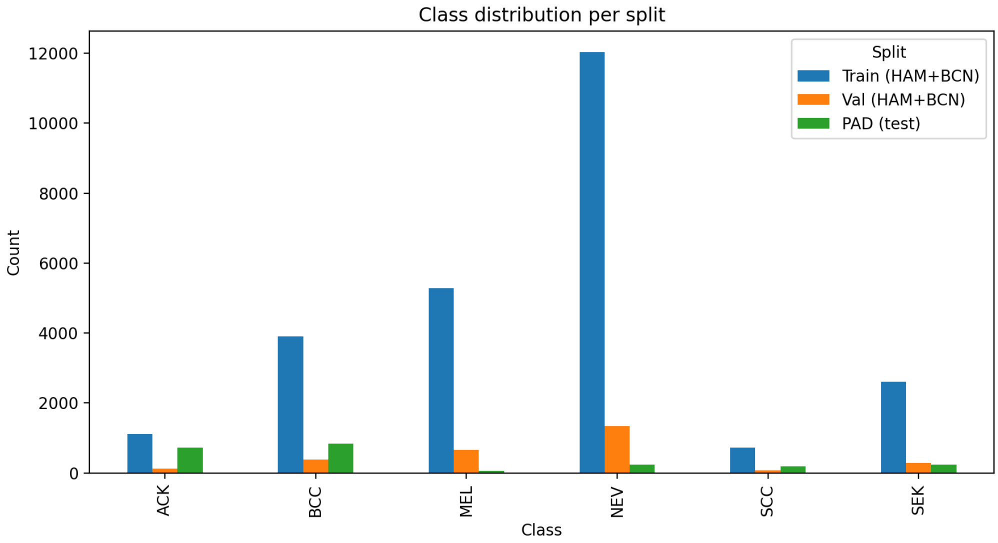


Figure 4.2: Class distribution across training (HAM+BCN), validation, and PAD test sets

Figure 4.2 demonstrates the distribution of classes by training and validation set of the source domain versus the test set of target-domain. As it is observed, the nevus (NEV) class, outweighs all the splits, and the malignant types such as the SCC and ACK are underrepresented. This imbalance in the classes is the driving force behind the decision to use macro-averaged F1-score as the key measure of evaluation in later experiments.

This unified mapping ensures that all models are trained to predict the same set of diagnostic categories regardless of dataset origin. It is also easier to make a direct comparison of a source-domain training and target-domain evaluation because the PAD-UFES-20 dataset is labeled in a similar manner. The images that are not in this unified set are not used in the experiments to ensure that all the labels are consistent and not add any contradiction to the learning experience.

4.2.3 Target Domain Dataset

The PAD-UFES-20 dataset is the target domain applied in this work as the evaluation one. PAD-UFES-20 provides clinical images taken with smartphone cameras, as opposed to using dermatoscopes, as in the case of the source-domain datasets. This difference in acquisition modality introduces a substantial domain shift that poses a significant challenge for automated classification models.

The PAD-UFES-20 images have a high variety of variations that are absent in dermoscopic datasets. These variations are uneven lighting, variation in color balance, appearances of background factors like hair or clothes, camera resolution, and variability of the lesion scale and orientation. Lesions in most instances take up a small fraction of the picture, making it even more challenging to classify.

In comparison to dermoscopic pictures whose quality is usually obtained by trained professionals and specialized equipment, smartphone pictures can be obtained by untrained individuals under uncontrolled conditions. Consequently, PAD-UFES-20 is more indicative of true-to-life application of mobile dermatology applications.

PAD-UFES-20 is used solely to test in this research. None of the images in this dataset are utilized in training, validation, hyperparameters tuning or model selection.This strict separation ensures that all reported results reflect genuine generalization rather than implicit adaptation to target-domain characteristics.

4.2.4 Domain Shift and Problem Formulation

One of the key issues that this study concerns is the issue of domain change between the dermoscopic image and the clinical images taken on a smartphone. Such change is based on underlying differences in imaging conditions such as acquisition equipment, light source, magnification as well as the degree of technical skill during image acquisition. Practically, these differences result in apparent differences in low-level visual features as well as high level appearance of lesions. The images of dermoscopes are obtained in a controlled clinical setting with the help of special equipment utilized by qualified individuals. Consequently, these images generally have a consistent lighting quality, even distribution of colors, sharp lines of lesions and low levels of interference at the back end. By contrast, smartphone pictures are frequently taken in free environments by non-professional cameras. These photos can have shadow, reflections, unequal light exposure, cluttered and messy background with hair or clothing, and much difference in the size and direction of the lesion. Automatic analysis is also complicated by the fact that in most instances, the lesion takes up a tiny part of the image.

The mismatch between the visual statistics of the target and source domains is intense due to such differences in the conditions of acquisition. Models that are only trained with dermoscopic images can thus learn features that perform well in controlled environments, but not strong to the variation brought about by smartphone photography. Such discrepancy is one of the main causes of such significant drop in performance of the traditional supervised models applied to the smartphone-acquired images directly. In this paper, the learning problem is formulated as a zero-shot cross-domain generalization problem. The training of models is conducted based only on the labeled cases of dermoscopic images that are based on the combination of the HAM10000 datasets with the BCN20000 datasets to create the source domain. The trained models are subsequently tested directly on the smart phone acquired data of the PAD-UFES-20 dataset, which is the target domain. Importantly, there is no usage of images in the target domain during any part of the training, validation, selection of hyperparameters or selection of a model. Consequently, the models have to fully depend on representations and decision making mechanisms obtained within the source domain, as well as inductive biases presented through the selection of the learning paradigm. There are no assumptions on the access to targets-domain data and no theoretical restrictions on domain-invariance. Rather, the goal is to empirically test the behaviors of the various learning strategies in a realistic and demanding deployment condition with extreme cross-domain shift. This formulation corresponds to practical limitations that are often faced in practice in medical applications in which labeled target-domain data collection might be non-practical because of cost, privacy issues, expert inaccessibility. This restrictive zero-shot condition ensures that the study concentrates on the true generalization ability as opposed to adaptation and they provide an objective view of the real-life scenario on the power of cross-device skin lesion classification.

4.2.5 Zero-Shot Constraints and Motivation

A number of restrictions are placed to guarantee the integrity of the zero-shot setting. The first one is that hyperparameter tuning and early stopping do not utilize target-domain images; performance on source-domain validation is used to make all model selection choices. Second, no type of test-time adaptation, calibration or pseudo-labeling is involved. Third, the source and target domains have the same set of preprocessing steps, and no implicit domain-specific adaptation is made. The stringent zero-shot approach that will be used in this work is driven by pragmatic factors in medical practice. In most real-world situations, labeling data of the target-domain is not easy because of the cost, privacy issues, or the absence of expert labeling. Additionally, it might be impossible to regularly train or fine-tune deployed models. This study analyzing models in the conditions of such constraints will offer an understanding of the constraints of the existing methods and will demonstrate the problems that need to be solved to achieve reliable cross-device classification of skin lesions.

4.2.6 Summary

This section has discussed the datasets that were to be used in the research and officially framed the problem of zero-shot learning that is to be solved by the suggested methodology. The experimental setup presents a serious and realistic domain shift by using two dermoscopic datasets as the source domain and assessing them on a smartphone-based target domain alone. The extreme division of source and target domain makes sure that all successive results will be the result of real generalization but not adaptation. This dataset design is the basis based on which the model architectures, learning strategies and experimental assessments presented below are constructed.

4.3 Model Specification

This section describes the model architectures and learning approaches employed in this study. The objective of this section is to explain what models are used, how they are structured, and why they are appropriate for evaluating zero-shot generalization under severe cross-domain shifts. Performance comparisons and quantitative analysis are deliberately deferred to Chapter:5 . To achieve clear methodology, model specification is designed in three sections, backbone network architectures, meta-learning model, and normal supervised learning baselines. This group indicates the factorised design of the suggested methodology, in which the representation of features and learning paradigm are considered discrete elements.

4.3.1 Backbone Network Architectures

Backbone network architectures serve as feature extractors that transform raw input images into high-level representations used for downstream classification. The quality, robustness and inductive bias of these representations is very important in the context of the zero-shot learning when there is a strong domain shift, as these affect the overall performance of generalization. Given that no data in the target domain are present in the training process, the backbone should not be learning features that depend on the change of acquisition conditions but rather features that are closely linked to the source domain. This paper compares various backbone designs based on classical and contemporary convolutional neural networks. Instead of critically examining all possible architectures, representative models are chosen to reflect diversity in architecture design, capacity and inductive bias. The choice allows the consideration of the interaction of various forms of feature representations with meta-learning strategies in case of severe cross-domain shifts. All backbone architectures are initialized with the publicly available prepared weights, which are usually trained on massive datasets of natural images. Pretraining gives a good initialisation that enhances convergence and stabilises training, especially in episodic meta-learning. Every backbone is incorporated into meta-learning and traditional supervised training pipelines where needed so that it is directly compared across the learning paradigms with the same representational constraints.

I. Convolutional Neural Network Backbones

The convolutional neural networks (CNNs) have traditionally been used in medical image analysis because of their very high inductive bias toward locality and translation invariance. Such properties are especially the most appropriate to dermoscopic images, in which discriminative cues tend to rely on localized texture, color change, and lesion boundary organization. Nevertheless, in extreme domain shift, these inductive biases can be beneficial and detrimental to generalization, respectively, based on the extent to which learned features are conditioned to features of source-domain imaging. In the analysis of these effects, this paper will involve CNN backbones that are based on various generations of architecture design starting with classical residual networks and modernized convolutional architecture based on transformer design principles.

ResNet

ResNet is a classical convolutional baseline representing earlier generations of deep CNN architectures. It is characterized by the use of residual connections, used to train deep networks, through which the vanishing gradient issues are mitigated. ResNet has grown to form a baseline in general computer vision and medical image analysis because of its simplicity, stability, and ubiquitous use. ResNet is used in the work more as a point of reference, as opposed to a state-of-the-art backbone. It enables the determination of the extent to which performance improvement is due to architectural modernization or learning strategy change. Although ResNet has shown to perform well in the dermoscopic datasets in cases of supervised learning, its dependence on somewhat rigorous convolutional inductive biases can restrict its performance in generalization in the case of extreme acquisition changes. By incorporating ResNet, clear baseline comparison will allow to separate the contribution of observed gains due to meta-learning, and due to stronger feature representations.

EfficientNet-B3

EfficientNet-B3 is chosen because it is widely applied in medical image tasks and has a fair balance between the representational and computational power. Efficient-

Net proposes a principled compound scaling, which currently scales network depth, width, and input resolution together to achieve better performance without overgrowing in parameters. EfficientNet variants have also demonstrated good performance in supervised learning where they have been used in the dermoscopic datasets and even taken the form of competitive baselines when it comes to classification of skin lesions. In terms of this research, EfficientNet-B3 is a high-capacity CNN that is trained using traditional supervised learning. Its addition allows this to be analyzed in terms of whether architectures that are created with efficient supervised optimization are robust to zero-shot domain shift when paired with meta-learning. In particular, EfficientNet-B3 allows investigation into whether increased representational capacity alone is sufficient to improve cross-domain generalization, or whether such capacity instead leads to stronger overfitting to source-domain characteristics.

ConvNeXt-Tiny

ConvNeXt-Tiny is an adapted convolutional architecture that uses design ideas inspired by vision transformers but uses the same inductive biases as convolutional networks. These design choices include larger kernel sizes, simplified stage configurations, and modified normalization strategies, resulting in architectures that more closely resemble transformer-like processing pipelines while preserving spatial locality. ConvNeXt-Tiny is chosen as the main CNN backbone in the given study because it is empirically strong and showed its ability to be used in episodic meta-learning systems. ConvNeXt-Tiny, in comparison with previous CNN architectures, has better representational flexibility, though with the same computational efficiency, which makes it especially efficient in repeated episodic training. ConvNeXt-Tiny is a conceptual linkage between conventional convolutional network and transformers models, methodologically speaking. Its addition gives the study an opportunity to determine whether modernized convolutional representations have a higher capacity to aid task-based learning and generalization in extreme domain shifts. ConvNeXt-Tiny is thus considered with respect to numerous meta-learning techniques in this publication. This large-scale assessment allows a thorough scrutiny of the nature of interaction between modern convolutional feature extractors and various meta-learning paradigms and whether such representations are always better than either classical CNNs or larger capacity supervised models on zero-shot evaluations.

II. Transformer-Based Backbones

Transformer-based models are gaining more and more popularity in computer vision because of their capability of capturing global dependencies by means of self-attention. In contrast to convolutional neural networks, which are locality-based and feature hierarchical representations, vision transformers are patch-based and directly substantiate long-range interaction between all parts of the image. This inductive bias is significantly different, and it is due to this architectural difference that robustness in the face of extreme domain shift may be possible. Global attention mechanisms can be useful in the context of dermoscopy-to-smartphone generalization to capture lesion-scale structure less susceptible to local changes in texture, lighting or background clutter. Nevertheless, the usefulness of transformers in a strict zero-shot regime remains an open question, since the large scale data or high strength pretraining is necessary to learn meaningful representations. To explore these trade-offs, both the supervised and self-supervised transformer-based backbones will be incorporated into the study, which vary in terms of training paradigm, inductive bias and representational properties.

DeiT-Small

DeiT-Small (Data-efficient Image Transformer) is a supervised vision transformer trained to minimize the data-steps required during the training of a transformer. It does so by better training approaches such as knowledge distillation by convolutional teacher model and perfect augmentation plans. DeiT-Small is therefore more stable in training and efficient than the previous variants of vision transformers. DeiT-Small will be considered as a representative of a supervised transformer backbone in this research. Its smaller architecture is computationally viable to episodic meta-learn, and it still has the characteristics of global attention found in transformers. Contrary to CNNs, DeiT-Small handles images as patches sequences, which can capture long-range dependencies, which can be useful in cases where the appearance of lesions is influenced by other image contextual artifacts like surrounding skin or background artifacts. The addition of DeiT-Small allows one to directly compare the convolutional with the transformer-based representation when all the training and evaluation conditions are the same. This comparison aids in finding out whether the global attention based representation offers more benefit than locality based convolutional features in the cases where a model is requested to generalize between acquisition devices and without target domain supervision.

DINOv2-Small

DINOv2-Small is a self-supervised vision transformer that is trained on large-scale unlabeled data on a contrastive learning objective. In contrast to supervised backbones, with DINOv2, representations are learned by ensuring consistency between multiple different views of the same image, as opposed to the classification objective being defined. The paradigm of this training facilitates the model to learn the general-purpose visual characteristics which can be less related to the dataset-specific labels or biases. This study is also able to investigate the hypothesis of stronger domain invariance in self-supervised representations compared to supervised representations when this study is conducted with DINOv2-Small as zero-shot evaluation testable representations. DINOv2 is unsupervised, which means that the features learned by it can be less sensitive to source-domain label distributions and acquisition properties. DINOv2-Small is mostly applied as a fixed feature extractor in meta-learnings in this work. This design allows the evaluation of whether the self-supervised representations with the help of task-level learning strategies can enhance the robustness to the cross-domain shift to the medical image classification.

Summary of Transformer Backbone Selection

Together, DeiT-Small and DINOv2-Small represent two distinct transformer-based design philosophies: supervised training with optimized data efficiency and self-supervised representation learning at scale. They allow studying the interaction of training paradigm and architectural bias with meta-learning when the domain shift is the extreme one. By these models with convolutional backbones, one can gain an idea of whether global attention mechanisms and self-supervised learning have benefits in the case of zero-shot medical imaging.

4.3.2 Meta-Learning Models

Meta-learning models form the core methodological component of this study. In contrast to traditional supervised learning, which tries to optimize the model parameters to achieve performance on a fixed set of tasks, meta-learning seeks to optimize the process of learning itself, i.e. by training models to be able to perform highly on a distribution of tasks. This task formulation can promote the learning of representations and decision strategies that can be extrapolated into a new distribution of training. Meta-learning is one of the possible mechanisms of enhancing robustness in the presence of severe domain shift that is used in this work. Meta-learning episodically trains models on the source domain and introduces structured variability to the models, potentially diminishing their dependence on source-specific indicators and enhancing their extrapolation to previously unseen domains. Five representative methods are considered to represent a wide variety of meta-learning paradigms: Prototypical Networks, Meta-Baseline, FEAT, MetaOptNet. These approaches vary in the way they do inference, task adaptation and exploitation of backbone representations.

Prototypical Networks

Prototypical Networks are a measure based meta-learning model which undertakes classification through calculation of class prototypes in an embedding space. In every episode, a prototype of each of the classes is constructed by averaging the embeddings of support samples of that class. The samples of the query are categorized according to their proximity to these prototypes. It is done by assuming that samples of the same class cluster tightly in the learned feature space. The Prototypical Networks are stable and computationally efficient to train, and this property suits them well to train large-scale episodic studies. Here, they are used as the main meta-learning baseline in the study and are extensively applied to compare backbone architectures in the same training conditions.

Meta-Baseline

Meta-Baseline extends Prototypical Networks by replacing distance-based inference with a learned classifier operating on feature embeddings. While training remains episodic, classification is performed using a parametric decision function, allowing for more flexible decision boundaries. This method bridges the gap between conventional supervised learning and meta-learning by combining task-based training with classifier-based inference. Meta-Baseline is included to assess whether increased decision flexibility improves robustness under severe domain shift.

FEAT

FEAT (Few-shot Embedding Adaptation with Transformer) suggests an embedding adaptation model, which modifies the support embeddings using attention-based operations and then calculates the prototypes. This adaptation of embeddings can be conditioned on the structure of an episode and can improve the separability of the classes. FEAT is a meta-learning approach that is more expressive, namely, it explicitly represents sample-sample relationships in an episode. In the present study FEAT is experimented to indicate whether these adaptive embedding transformations increase zero-shot generalization in cases where there is a big difference between the distribution of features between the source and the target domain.

MetaOptNet

MetaOptNet develops a convex optimization model of classification by training a support vector machine on the embeddings generated by the backbone net. Instead of using fixed distance metrics or learned classifiers, MetaOptNet uses an episode-specific decision boundary to optimize a task-specific decision boundary. The method can provide a more flexible separation of classes in embedding space, and can explicitly control margin-based classification behaviour. It is accompanied by MetaOptNet to investigate the idea of better performance by episode-specific decision boundaries in zero-shot performance during a drastic domain shift.

4.3.3 Normal Supervised Learning Baselines

In order to put the performance of meta learning approaches in perspective, traditional supervised learning baselines are also incorporated in the study. These baselines utilize backbone architectures that are the same as those in the meta-learning experiments but are trained by standard mini-batch optimization with cross-entropy loss instead of episodic sampling. These baselines are not aimed at competing with meta-learning models on an absolute basis but to create a reference point that is controlled in the sense that it isolates the effect of the learning paradigm. The study allows to directly compare conventional supervised learning and meta-learning with the same representational restrictions by keeping the backbone architecture constant and only changing the training strategy. In normal supervised training, the models are trained to minimize the error of classification on each sample of the source domain. Although this method has worked well when there are equal training and testing conditions, it makes models to exploit source specific visual cues which might not be transferable to other acquisition devices. Consequently, trained models are frequently characterized by steep performance drop in the case of severe domain shift evaluation. By adding supervised baselines, the study will be capable of determining whether meta-learning performs worse than traditional training in the context of using an unseen target domain. Instead of comparing the peak source-domain accuracy, the comparison focuses on robustness and transferability, which are the focus of the research objectives. Because of computational limitations, normal supervised baselines are only applied to a set of backbone architectures that have been specifically chosen and not all the models that were used in the meta-learning experiments. This methodical bias is not by chance and it is consistent with the methodological focus of the research. The backbones are selected to reflect the various families of architectures and capacities, such that one cannot make conclusions only related to one type of model. Notably, trained and evaluated supervised baselines are trained and evaluated under the same preprocessing, data splits and evaluation protocol performed on meta-learning models. This consistency means that the differences in performance in the target domain can be related to the learning paradigm as opposed to confounding variables including data leaks or variations in training conditions.

4.3.4 Summary of Model Specification

The model architectures and learning strategies used in the study have been defined in this section. The collection of backbone networks is varied to include variation in architectural design, representational capacity and inductive bias. Such backbones are coupled with several meta-learning paradigms, which exemplify different methods of task-level learning and adaptation, which are metric-based, classifier-based, optimization-based, and gradient-based. Besides meta-learning models, the best normal supervised baselines are also provided which offer a controlled comparison and isolate the role of episodic learning. These elements constitute a systematic experimental design which allows a systematic study of the effect of representation choice and learning strategy on zero-shot generalization in strong cross-domain shifts. The proposed model specification can be used to make fair and interpretable comparisons between methods by ensuring that training conditions and a set of zero-shot evaluation rules are adhered to. The resulting framework is not just meant to quantify performance but to uncover trends in robustness and failure modes that will be used in the real world implementation of the skin lesion classification systems on the acquisition devices. In the next section, the models are trained and assessed using a common experimental procedure, which guarantees consistency of reproducibility and methodological consistency of all experiments.

4.4 Training and Evaluation Strategy

In this study, the training procedures and evaluation protocol are developed in this section. This is aimed at establishing explicit training of the models under homogeneous conditions and the evaluation of the performance under a strict zero-shot environment. All the methodological decisions are made in accordance with the design principles discussed earlier in this chapter to make sure that no data of the target domain affect the training, validation, and model selection.

4.4.1 Data Preparation and Training Procedures

Both source and target domain images are processed through one preprocessing pipeline to provide consistency and prevent unspoken domain-specific adaptation. The images get resized to a fixed resolution that is supported by the backbone architectures of interest and standard normalization is done using statistics that are aligned with the relevant pretrained models. The merged datasets of HAM10000 and BCN20000 are divided into a training set and a validation set when it comes to the source domain. The model optimization is done on the training split and the model is then selected and early stopped in case of validation split. Both splits have no information on any images of the target domain. PAD-UFES-20 dataset is not used at all, except as a held-out test set, and never learned on when training, validating, tuning hyperparameters or selecting checkpoints. This hard distinction is made so that any reported results are actual zero-shot generalization. It uses two training strategies: normal supervised training and meta-learning based episodic training.

Models (when being optimized by mini-batch stochastic gradient descent with cross-entropy loss) are also optimized in normal supervised training. This method is used as a benchmark and it is common in the practice of medical image classification. The source domain episodic sampling is used to train the meta-learning models. In every episode, there is a simulated classification task, which involves splitting the samples into support and query sets. The model is also tailored to be effective when such tasks are distributed as opposed to single samples. In gradient-based methods, inner-loop adaptation is also limited by the algorithm design with no control over it and in non-gradient-based methods, task-specific inference is made by metric-based, classifier-based, or optimization-based mechanisms.

4.4.2 Optimization and Evaluation Protocol

Models of this paper are all optimized with stochastic gradient-based optimization and only trained on source-domain data. The entire process of making optimization choices is done without having samples of target domains in order to ensure the integrity of the zero-shot analysis. The choice of key training hyperparameters is also based on initial experiments on the source-domain validation set. They are learning rate, training epochs, and batch size in normal supervised training and episodic configuration parameters in meta-learning methods, including: the number of classes in each episode (ways), the number of support samples per class (shots), and the number of query samples. One of these hyperparameters is stored permanently after selection between similar experiments. Only necessary minor changes are done to hyperparameters to guarantee stable optimization. Particularly, learning rates are varied among backbone families to support convergence behavior variations between convolutional and transformer-based architectures. Moreover, the number of training epochs is further augmented with higher-capacity backbones so that there is adequate convergence in the source domain. In the case of meta-learning algorithms, episodic batch-sized is scaled down to fit the memory limits of the standard GPUs without changing the episodic formatting. Target-domain performance is not involved in the determination of changes. To be fair, experiments with the same backbone and learning paradigm have the same hyperparameter settings. This will guarantee that the differences in performance can be explained by the learning strategy or representation as opposed to the differences in tuning.

4.4.3 Evaluation Metrics and Experimental Design Rationale

The measure of performance is macro-averaged F1-score, which attaches the same value to every diagnostic class and fits well an imbalanced medical dataset with classes. The overall classification accuracy is given as the secondary measure to offer extra awareness but is not utilized as the main basis of comparison. The design of the experiment is more interpretative and methodological than exhaustive. Backbone architectures are initially tested within a general meta-learning framework that investigates the robustness of representation. The backbones are then selected in order to compare various meta-learning approaches to determine the variation in the learning paradigms. Normal controlled baselines are added to give contextual performance of meta-learning when limited to the same architectural constraints. This designed assessment plan will enable the study to be able to isolate the influence of representation and learning paradigm, and also make the computationally feasible. The methodology achieves this by ensuring that the variations in observed performance can be explained by an architectural choice or training strategy rather than by uncontrolled experimental variation by fixing the evaluation protocol and varying only one factor at a time. Notably, the study design does not make conclusions based on individual performance figures. Rather, the focus is put on stable patterns across models, including relative robustness in domain shift and relative degradation of supervised and meta-learning methods. This methodology will work in line with the main goal of the study, namely to measure generalization behavior instead of maximizing peak accuracy on the source domain. The combination of the chosen evaluation measures and the designed controlled experiment will offer a reasonable and interpretable framework of comparing the models in a strict zero-shot environment. This framework assists in carrying out the meaningful analysis of the impact of various architectural and learning decisions on the robustness in the case of models trained on dermoscopic images when they are applied to smartphone-obtained clinical images.

Chapter 5

Result Analysis

5.1 Performance Evaluation

In this section, the results of the experiment of the proposed zero-shot skin lesion classification framework are presented. The main goal of this assessment is to determine the performance of various backbone structures and learning paradigms in the scenario where the performance is under heavy cross-domain shift, especially when using models that are trained on dermoscopic images in direct application to smartphone-based images without target-domain adaptation. Any findings that are presented in this section are in line with the rigor assessment guideline as outlined in Chapter 4. Evaluation is done on the seen-domain validation set, and the unseen target-domain test set (PAD-UFES-20). The main evaluation measure is the macro-averaged F1-score, and the overall classification accuracy is presented as a supporting one.

5.1.1 Evaluation Setup

To assess the model performance, there are two conditions that are taken to analyze all the experiments:

- Seen-domain evaluation, conducted on the source-domain validation split derived from the combined HAM10000 and BCN20000 datasets.
- Unseen-domain evaluation, conducted exclusively on the PAD-UFES-20 dataset, which consists of smartphone-acquired clinical images.

No samples from the target domain are used during training, validation, hyperparameter tuning, or model selection. This strict separation ensures that all unseen-domain results reflect genuine zero-shot generalization rather than implicit adaptation. Both Macro-F1 and accuracy are reported in each experiment to get a fair picture of both class-sensitive and overall performance. Nevertheless, Macro-F1 is considered as the main measure because of the imbalance in classes that are presented in skin lesion samples and the clinical significance of minor classes.

5.1.2 Normal Supervised Learning Performance

This subsection presents the performance of standard supervised learning models trained on the source domain using mini-batch optimization and cross-entropy loss.

Five representative backbone architectures are evaluated under this setting: ResNet, ConvNeXt-Tiny, EfficientNet-B3, DeiT-Small, and DINOv2-Small.

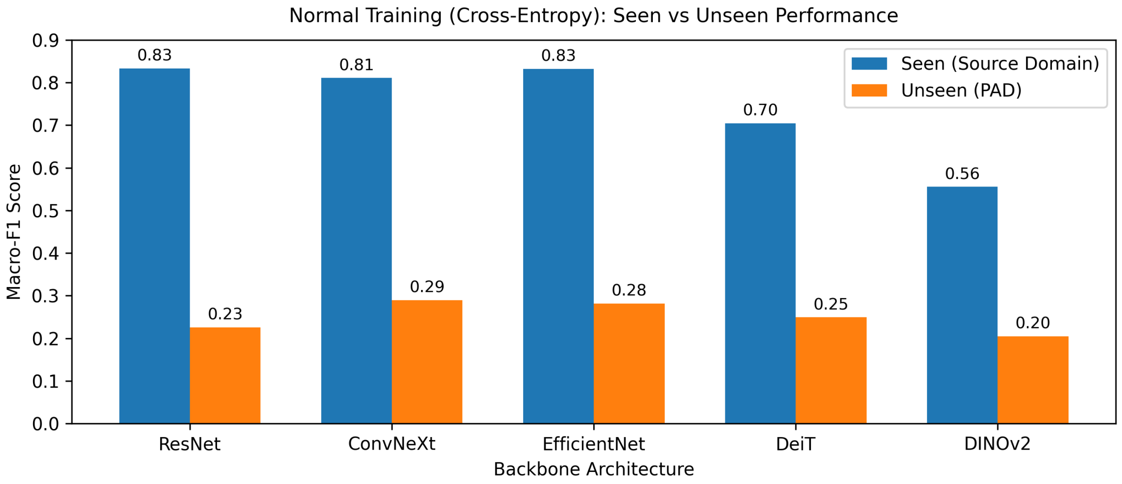


Figure 5.1: Normal training seen vs unseen performance

Figure 5.1 summarizes the seen and unseen performance of these models. Normal supervised training among all backbones provides good performance on the seen source domain, Macro-F1 scores of mostly between 0.70 and 0.83 and accuracy of above 0.78 in most of the cases. These findings correlate with the previous studies of dermoscopic datasets and prove that the chosen architectures can be trained to learn discriminative features with controlled conditions of imaging. However, on the unseen PAD-UFES-20 data, all backbones plummet in performance. The Unseen Macro-F1 scores are brought down to 0.20-0.29 range, and the corresponding decline in accuracy is made. This degradation is seen in a uniform manner in convolutional, transformer-based, and self-supervised backbones. Notably, even architectures with high seen-domain performance, including EfficientNet-B3 and ConvNeXt-Tiny, do not show good performance on images captured on a smartphone. This action emphasizes how harsh dermoscopy-smartphone domain shift can be and shows that intense source-domain optimization cannot ensure effective cross-device generalization. These results establish a critical baseline observation: high performance on dermoscopic images does not translate to reliable performance on smartphone images under a zero-shot setting.

Table 5.1: Seen vs Unseen Performance under Normal Supervised Training

Backbone	**Seen F1**	**Unseen F1**	**F1 Drop (%)**	**Seen Acc**	**Unseen Acc**	**Acc Drop (%)**
ResNet	0.83	0.23	72.3%	0.87	0.28	67.8%
ConvNeXt-Tiny	0.81	0.29	64.2%	0.87	0.35	59.8%
EfficientNet-B3	0.83	0.28	66.3%	0.89	0.33	62.9%
DeiT-Small	0.70	0.25	64.3%	0.79	0.29	63.3%
DINOv2-Small	0.56	0.20	64.3%	0.70	0.29	58.6%

Drop (%) denotes the relative decrease when transferring from the seen (source) domain to the unseen PAD-UFES-20 domain.

Table 5.1 also highlights the substantial performance degradation when models trained under normal supervised learning are evaluated on the unseen PAD-UFES-20 domain. The backbones of all reported high seen-domain performance, however, their unseen Macro-F1 and accuracy experience a dramatic loss (58-72%), that is, domain shift is severe between dermoscopic and smartphone images. This reaffirms that the common supervised training is ineffective in generalizing across devices of acquisition.

5.1.3 Meta-Learning with Prototypical Networks Across Backbones

Prototypical Networks are trained on the source domain using episodic training to assess the ability of each of the five backbone architectures. To evaluate whether meta-learning can improve robustness under severe domain shift. This allows a controlled comparison between normal supervised learning and metric-based meta-learning under identical architectural constraints.

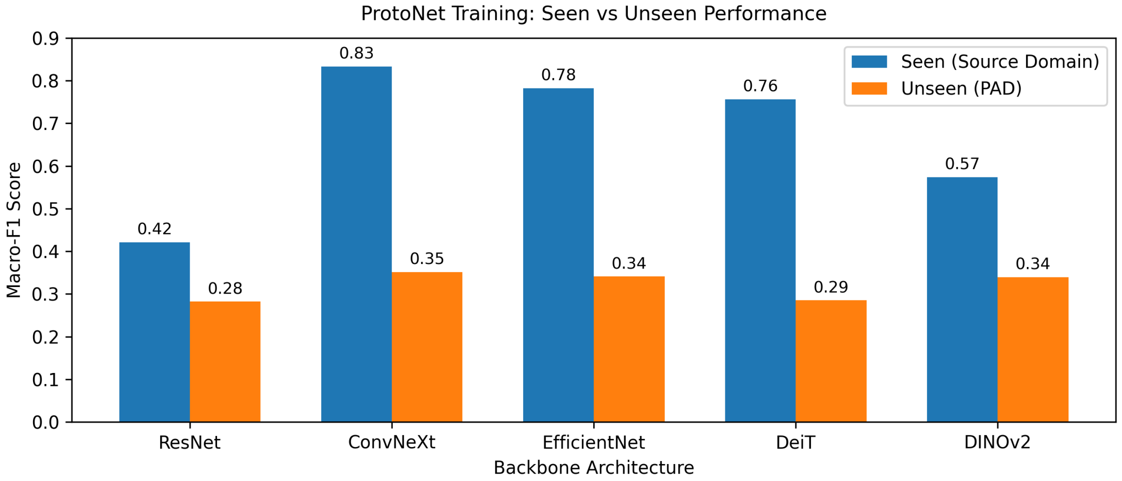


Figure 5.2: Protonet training seen vs unseen performance

Figure 5.2 reports the seen and unseen performance of Prototypical Networks for each backbone.

Across all backbones, Prototypical Networks produce consistent improvements in unseen-domain Macro-F1 compared to their normal supervised counterparts. The magnitude of improvement varies by architecture but is observed universally:

- ResNet demonstrates the growth of unseen Macro-F1 between 0.23 and 0.28.
- ConvNeXt-Tiny improves from 0.29 to 0.35.
- EfficientNet-B3 improves from 0.28 to 0.34.
- DeiT-Small improves from 0.25 to 0.29.
- DINOv2-Small has the greatest improvement and it has grown by 0.20 to 0.34.

Table 5.2: Seen vs Unseen Performance under ProtoNet Training

Backbone	**Seen F1**	**Unseen F1**	**F1 Drop (%)**	**Seen Acc**	**Unseen Acc**	**Acc Drop (%)**
ResNet	0.42	0.28	33.3%	0.54	0.34	37.0%
ConvNeXt-Tiny	0.83	0.35	57.8%	0.88	0.43	51.1%
EfficientNet-B3	0.78	0.34	56.4%	0.83	0.40	51.8%
DeiT-Small	0.76	0.29	61.8%	0.80	0.32	60.0%
DINOv2-Small	0.57	0.34	40.4%	0.66	0.40	39.4%

Drop (%) denotes the relative decrease when transferring from the seen (source) domain to the unseen PAD-UFES-20 domain under ProtoNet training.

Table 5.2 shows that ProtoNet training substantially mitigates the performance drop observed under normal supervision when evaluated on the unseen PAD-UFES-20 domain. In comparison to Table 5.1, the relative loss of both Macro-F1 and accuracy is more consistently less across all backbones and in specific cases, ConvNeXt-Tiny and DINOv2-Small have significant gains. This implies that episodic meta-learning allows the models to acquire more transferable representations, which enhances cross-device domain shift resistance.

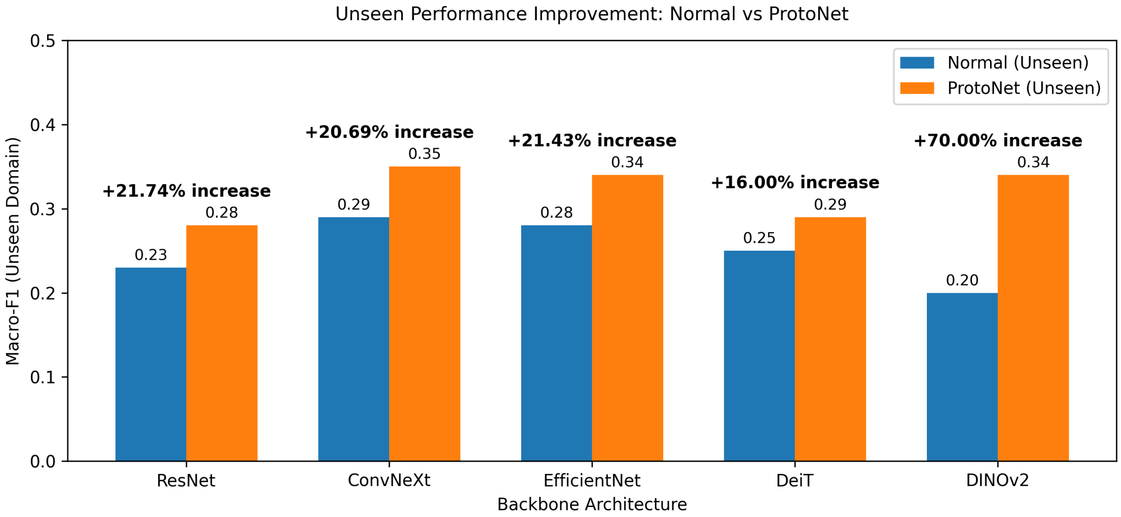


Figure 5.3: Cross domain performance improvement with meta learning

These findings suggest that episodic training and prototype based inference reduce performance decay across cross-domain shifts between all tested architectures. Compared to this, the observed behaviour in seen-domain performance under Prototypical Networks is more diverse.For some backbones, such as ConvNeXt-Tiny and DeiT-Small, seen-domain performance remains comparable to or slightly higher than that of normal supervised training. In other cases, especially that of ResNet, the performance in seen domains significantly reduces. This trade-off captures the underlying distinction between the supervised optimization and episodic learning where the goal is not to maximize the accuracy of the source domain after but to enhance the level of task generalization. Notably, the performance decrease of the seen-domain does not reject the advantages witnessed on the unseen domain. Instead, it suggests that meta-learning prioritizes representations and decision mechanisms that are less specialized to source-domain statistics, resulting in improved robustness when deployed on unseen devices.

Table 5.3: Unseen Domain Performance Improvement (Normal Training → ProtoNet)

Backbone	**Normal Unseen F1**	**ProtoNet Unseen F1**	**Δ F1**	**F1 Gain (%)**
ResNet	0.23	0.28	+0.05	21.74%
ConvNeXt-Tiny	0.29	0.35	+0.06	20.69%
EfficientNet-B3	0.28	0.34	+0.06	21.43%
DeiT-Small	0.25	0.29	+0.04	16.00%
DINOv2-Small	0.21	0.34	+0.13	61.90%

Δ F1 denotes the absolute improvement in unseen-domain Macro-F1, while F1 Gain (%) indicates the relative percentage improvement achieved by ProtoNet over normal supervised training.

Table 5.3 is a quantitative comparison of the performance of ProtoNet and normal supervised training on unseen-domain performance. Across all backbones, ProtoNet consistently increases Macro-F1 on the PAD-UFES-20 dataset, with gains ranging from 16.0% to 61.9%. The largest relative improvement is observed for DINOv2-Small, indicating that meta-learning is particularly effective for backbones trained with limited task-specific supervision. Such findings also verify that episodic training enhances cross-domain generalization without the need to have target-domain data.

5.1.4 Backbone-wise Comparison Under Meta-Learning

ConvNeXt-Tiny achieves the strongest unseen-domain performance under the evaluated zero-shot setting among the assessed ones when used in meta-learning. ConvNeXt-Tiny, combined with Prototypical Networks, has the highest score in unseen-domain Macro-F1 of all backbones, and is also effective in seen-domain. EfficientNet-B3 and DINOv2-Small also reflect significant improvements with regards to Prototypical Networks, but their overall unseen performance is a bit lower than ConvNeXt-Tiny. DeiT-Small demonstrates the smaller gains whereas ResNet, in spite of having a meta-learning advantage, is constrained by the age of its architecture. On the basis of these results, ConvNeXt-Tiny is chosen as the main backbone to be used in the subsequent analysis of meta-learning. Such choice is made by empirical performance on the unseen domain as opposed to source-domain accuracy, which is consistent with the very aim to enhance zero-shot generalization in the presence of device-level domain shift.

Table 5.4: Seen vs Unseen Performance (Macro-F1 and Accuracy) Across Backbones and Learning Paradigms

Model	Seen F1	Unseen F1	Seen Acc	Unseen Acc
ResNet (Normal)	0.83	0.23	0.87	0.28
ResNet (ProtoNet)	0.42	0.28	0.54	0.34
ConvNeXt (Normal)	0.81	0.29	0.87	0.35
ConvNeXt (ProtoNet)	0.83	0.35	0.88	0.43
ConvNeXt (Meta-Baseline)	0.76	0.36	0.80	0.41
ConvNeXt (FEAT)	0.79	0.34	0.83	0.41
ConvNeXt (MetaOptNet)	0.80	0.33	0.84	0.42
EfficientNet (Normal)	0.83	0.28	0.89	0.33
EfficientNet (ProtoNet)	0.78	0.34	0.83	0.40
DeiT-Small (Normal)	0.70	0.25	0.79	0.29
DeiT-Small (ProtoNet)	0.76	0.29	0.80	0.32
DINOv2 (Normal)	0.56	0.21	0.70	0.29
DINOv2 (ProtoNet)	0.57	0.34	0.66	0.40

Seen = source-domain validation performance; Unseen = PAD-UFES-20 zero-shot performance.

5.1.5 Extended Meta-Learning Evaluation on ConvNeXt-Tiny

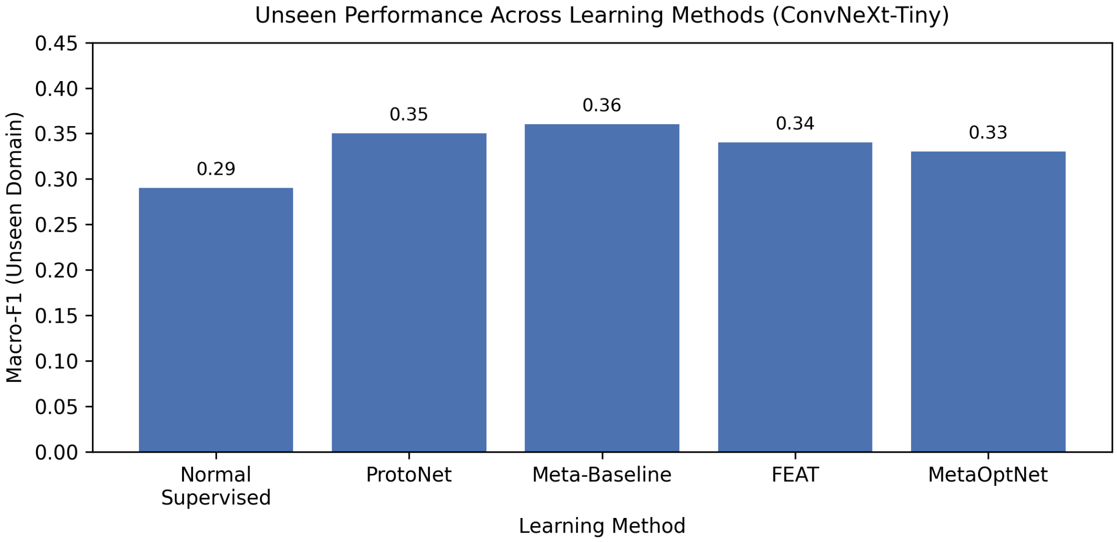


Figure 5.4: Different meta models on Convnext-tiny

Further variations in the effects of various meta-learning paradigms are considered by assessing further meta-learning techniques based on the ConvNeXt-Tiny backbone. These are Meta-Baseline, FEAT, and MetaOptNet as well as Prototypical Networks. Findings also show that Meta-Baselines improvement over Prototypical Networks is modest, the trend is consistent and observed under identical evaluation conditions has the best unseen-domain Macro-F1 score (0.36), slightly beating Prototypical Networks (0.35). FEAT and MetaOptNet also enhance the unseen performance over regular supervised training, but their increases are smaller and less steady. The results indicate that even though Prototypical Networks receive most of the boosts in robustness of episodic learning, incremental boosts can be made by more expressive meta-learning approaches. Yet, the size of these gains is small in comparison to the first boost in the case of switching between normal supervised learning and episodic meta-learning.

5.1.6 Summary of Performance Evaluation

In summary, the findings found in this section show that:

- Normal supervised models can perform well on the source space but have serious performance breakdown on unseen images on smartphones.
- Prototypical Networks consistently enhance performance in unseen-domain in every of the assessed backbones.
- ConvNeXt-Tiny is the most robust and stable when used in meta-learning.
- More elaborated meta-learning techniques can give slight advances over Prototypical Networks but fail to change the overall tendencies.

These results give undisputable empirical support that meta-learning decreases the decline in performance during extreme cross-domain transfers and encourages further investigation in the subsequent sections.

5.2 Analysis of Design Solutions

Here, the review of the experimental results contained in the Section 5.1 is conducted regarding the methodological design options used in the study. Instead of analyzing the results with specific numbers, the objective of the analysis is to describe the trend of the observed performance on the basis of analyzing the combination of backbone architectures, learning paradigms, and the environment of zero-shot deployment. The three main aspects of the proposed design that the discussion will be organized around include:

- The effect of episodic meta-learning compared to normal supervised training.
- The influence of backbone architecture on cross-device generalization.
- The behavior of various meta-learning strategies in the case of severe domain shift.

5.2.1 Impact of Episodic Meta-Learning on Zero-Shot Generalization

One of the key observations on the findings of the experiment is that meta-learning is always associated with an improvement in zero-shot performance on unseen smartphone images on all considered backbones. Although normal supervised models have a good performance on the seen source domain, they have a significant degradation on PAD-UFES-20 dataset. This degradation indicates excessive dependence on the features of dermoscopic images, which are not replicated by the images taken through smartphones.

On the other hand, Prototypical Networks perform better on the unseen domain of all the backbones. This is improved even though there are no target-domain samples provided during training and this observation supports the value of episodic learning in promoting transferable representations.

A possible solution to this phenomenon is that episodic training is task oriented. Meta-learning discourages overfitting to overall source-domain statistics and encourages less sensitive-to-acquisition artifact class-level discrimination by repeatedly solving small tasks of classification sampled out of the source domain. Consequently, the learned feature space is more robust to illumination, scale, and background material-changes-factors that prevail in the dermoscopy-to-smartphone shift in domain. Notably, the unseen performance increase is frequently accompanied by a decrease in the performance seen domain of some backbones. This trade-off aligns with previous reports in the literature of domain generalization and is indicative of the changing optimization objectives between the focus of maximizing the accuracy of the source-domain and the focus of maximizing task-level generalization.

5.2.2 Backbone Architecture and Representation Robustness

The findings also suggest that backbone architecture is an important factor that influences the effectiveness of supervised and meta-learning methods in cross-domain shift of device.

ConvNeXt-Tiny has the highest and most consistent performance on the unseen domain on the normal training and meta-learning of the backbones evaluated. Its contemporary convolutional structure is a combination of significant bias in locality with architectural elements of transformer models, and it is capable of capturing discriminative lesion information but resists noise and change in the background.

Episodic training is also advantageous to EfficientNet-B3, albeit with a larger capacity, that high capacity does not seem to provide a diminishing marginal benefit in the strictest instance of zero-shot scenarios. Although it is doing well on the source domain, its invisible performance is a bit worse than that of ConvNeXt-Tiny. This is an indication that larger model capacity does not necessarily result in greater robustness in domain shift.

Transformer backbones exhibit more diverse behaviour. DeiT-Small achieves moderate improvements with the help of meta-learning, yet the absolute unseen performance of the former is still worse than the one of convolutional models. This can be explained by the fact that it uses patch-based representations which are more susceptible to background distractions and changes in scale that are found in smartphone images.

The pattern observed in DINOv2-Small is unique: the performance of the model on the unseen domain is rather poor at its regular supervised performance, but it increases massively when combined with Prototypical Networks. This implies that self-supervised representations, not necessarily directly consistent with classification goals, can represent more domain-invariant features, which can be task-structured by metric-based inference.

5.2.3 Comparison of Meta-Learning Strategies

The long-term analysis of ConvNeXt-Tiny allows concluding that various meta-learning approaches can offer a higher or lower level of improvement over Prototypical Networks. The largest unseen-domain performance is by Meta-Baseline, which suggests that classifier-based meta-learning is able to provide marginal gains over simple distance-based inference when used on strong backbone representations.
FEAT and MetaOptNet are also more robust compared to regular supervised training and are not always better compared to Prototypical Networks. This implies that when the meta-learning head becomes more complex, there are no guarantees of corresponding improvement in generalization performance, under extreme domain shift.
In general, these results suggest that the main robustness gains are produced by episodic training as such, and not by the particular selection of meta-learning head. More expressive meta-learning procedures offer refinements to an existing performance pattern, but do not change the performance patterns fundamentally.

5.2.4 Summary of Design Analysis

To conclude, the discussion, conducted in this section, reveals the following main points:

- In the case of violent cross-domain shift, episode meta-learning achieves a consistent performance decrease.
- Backbone architecture has a great impact on zero-shot robustness, and more stable convolutional models have been observed to perform better.
- Meta-learning is advantageous to self-supervised representations, although it performs worse in supervised training.
- More complex meta-learning schemes provide only slight benefits over Prototypical Networks but they do not outcompete the simpler methods.

These observations help to justify the design decisions made in this study and offer a basis to the experimental refinements and comparisons as discussed in the following sections.

5.3 Final Design Adjustments

This part records the most important design refinements and implementation decisions as they occurred in the process of experimentation. Instead of reporting new performance outcomes, the aim of this part is to show how the experimental design changed after the observed behaviors and practical limitations, and how these changes impacted the final assessment outcomes. These adjustments are iterative, which is realistic research practice, especially where deep learning experiments are performed under constrained computational resources and severe zero-shot conditions.

5.3.1 Motivation for Design Refinement

Preliminary experiments with basic supervised learning demonstrated that there was a significant performance difference between the target-domain and the source-domain assessment. Although models had high Macro-F1 scores on dermoscopic validation data, the results were significantly low when the models were used on images obtained on smart phones.

This practice encouraged the transition of source-optimized training to learning paradigms that focus on robustness and transferability. Episodic training, and to a large extent, meta-learning, was thus embraced as the dominant approach to cross-domain shift when it comes to cross-device training.

As experiments went on, some refinements were made to help stabilize training, enhance generalization and to make sure that training was fair across models.

5.3.2 Episodic Training Configuration

A key design adjustment involved the formulation of episodic tasks that were applied in meta-learning. The episodes were designed so that there was a balanced representation of classes and samples such that no one training task was dominated by the majority classes.

Experiments were kept constant in terms of the number of classes per episode, the number of support samples and the number of query samples. A 6-way episodic training strategy was adopted instead of the commonly used 5-way setting to address class imbalance within the dataset. In the 5-way configuration, the dominant class (NEV) consistently appeared in every episode while one minority class was omitted, introducing a systematic bias. The 6-way formulation ensured uniform class inclusion across episodes.This nominalized episodic structure also allowed showing that variation in performance was due to the backbone architecture or learning paradigm and not the difference in the difficulty of the task.

The study minimized confounding by the establishment of regular episodic configurations which could be brought about by uneven sampling of tasks.

5.3.3 Normalization Strategy

In preliminary experiments, it was noted that batch normalization layers were vulnerable to variations in batch composition and data distribution, specifically in episodic training and a well-augmented environment. This sensitivity resulted in unreliable training dynamics and validation performance of some backbones.

To solve this problem, group normalization was embraced to undertake selected experiments only on training dataset. As opposed to batch normalization, group normalization relies not on batch-based statistics and thus can be used with different batch sizes and episodic settings. The modification helped to achieve a better stability of training and less variability of meta-learning performance, especially with more complex backbone architectures.

5.3.4 Data Augmentation and Robustness-Oriented Transformations

In order to further mimic variability that exists in images obtained through smartphones, further augmentation approaches were added to the training pipeline. These extensions were meant to add a controlled variability of color, illumination and compression artifacts which are similar to the standard traits of mobile imaging.

Instead of making augmentations domain-specific, these transformations were made in similar ways across source-domain data. This design decision ensured that no implicit information about the target domain was added during training and did not interfere with the integrity of the zero-shot setting.

The addition of robustness-based augmentations led to the higher unseen-domain performance, especially in the case of episodic training.

5.3.5 Backbone Selection and Experimental Focus

As results accumulated across backbones and learning paradigms it became clear that not every architecture reacted equally to meta-learning. ConvNeXt-Tiny was able to achieve good and consistently high performance on the unseen domain when using Prototypical Networks, and the performance when using seen-domain is also competitive.

On the basis of these findings, ConvNeXt-Tiny was chosen as the main framework on which to conduct further meta-learning-based evaluation. This choice made it possible to analyze more meta-learning strategies without being prohibitively expensive computationally.

Other backbones have been kept like ResNet and EfficientNet-B3 still used as comparative baselines to put into perspective the observed trends and make sure that conclusions did not depend on a specific architecture.

5.3.6 Handling of Practical Constraints

Due to computational and time constraints, not all experiments could be repeated across multiple random seeds or hyperparameter configurations. Rather, experimental design was aimed at breadth of comparison and consistency of trends across models. Where there was a minor difference in performance, focus was put on relative trends, as opposed to absolute. This method is in accordance with the main goal of the study, that is to determine robustness in terms of domain shift as opposed to maximization of source-domain accuracy.

Such limitations are specifically identified, and they are also addressed later in the relation to the interpretation of results in following sections.

5.3.7 Summary of Design Adjustments

Overall, the ultimate experimental design is the result of several design choices that are intended to make it more robust and interpretable within the rigid zero-shot conditions. Major advancements are standardized episodic training settings, normalization strategy choice, robustness-based data augmentation, and focused backbone choice.

These adaptations, combined, were the guarantee that the reported results are not the results of the unsteady training or arbitrary experimentation, but the result of the controlled and methodologically based evaluation process.

5.4 Statistical Analysis

This part talks about the statistical issues surrounding the experimental findings reported in this chapter. It is not aimed at offering exhaustive statistical testing, but only to explain the way the variability and reliability are addressed under the practical limitations of the study.

Most experiments in this work are performed with a single random seed per configuration because of the computational constraints. Consequently, the standard significance testing based on statistical tests on several independent runs are not conducted. Such a decision puts the topic of broader methodological coverage, including the assessment of a variety of backbones and learning paradigms, over variance estimation, which is a natural and acceptable compromise in undergraduate-level research.

Rather than a repetition, the reliability of results is determined based on the similarity of observed trends in various architectures and learning strategies. Specifically, the key message of this research is that meta-learning enhances zero-shot performance on unseen smartphone images can be justified by systematic changes of unseen-domain Macro-F1 across all tested backbones as a result of switching to episodic meta-learning. It is a fact that such improvement is not seen only in cases, but is seen consistently, which is more evidence than one statistically significant result on one model.

Moreover, Macro-averaged F1-score is employed as the main evaluation measure to decrease the sensitivity to the class imbalance to guarantee that the improvement is not dictated only by the dominant classes. It reports accuracy as a secondary measure to give more information and its trends are observed to be consistent with those of Macro-F1 in all experiments. Although the results could be further reinforced by more detailed statistical analysis, e.g. by multi-seed analysis or hypothesis testing, the overall directionality of findings across the backbones and the meta-learning strategies supports the notion that the identified improvements are not a result of random fluctuations but there is actual variance in the behavior of generalization.

The limitations imposed by single-seed evaluation are acknowledged and discussed further in the final discussion section. However, the statistical data provided in this paper is enough to justify the conclusions of the research within the limits and restrictions of the study.

5.5 Comparisons and Relationships

In this section, the relationships were studied between various variables of the experiment and placed the results of this work in the framework of the previous research on the problem of classifying skin lesions, domain generalization, and meta-learning. Instead of minding specific numerical outcomes, the analysis dwells on comparative patterns and structural associations that manifest themselves in systematic assessment.

5.5.1 Relationship Between Seen and Unseen Performance

One of the main relationships that occurred in all the experiments is the weak correlation between seen domain and unseen domain performance. Models with large Macro-F1 scores on the source-domain validation set are not always good performers on the target-domain test set. This is always evident with all the backbone architectures and learning paradigms that are tested in this study.

Normal supervised models, especially, are strongly optimized which occurs on the source domain and degrades significantly when tested on smartphone-acquired images. It means that source-domain accuracy is not a reliable predictor of deployment robustness in the case of extreme domain change.

However, meta-learning methods, in particular Prototypical Networks, were more likely to minimize this gap. As much as episodic training can reduce or even stabilize source-domain performance, it always enhances unseen-domain performance. This negative correlation highlights one of the key study results: effective generalization in the case of the domain shift involves compromising a certain level of source-domain specialization.

5.5.2 Backbone Capacity Versus Robustness

The other relationship which is important is the interaction between backbone capacity and zero-shot generalization. Although bigger or more complicated architectures may be fast on the source domain, it does not necessarily map to fast on the invisible target domain.

EfficientNet-B3, such as, with efficient supervised training, has good seen-domain performance, but with meta-learning, it shows only mediocre results compared to ConvNeXt-Tiny. This implies that architectural capacity is not enough to be robust in case of domain shift.

ConvNeXt-Tiny exhibits a positive representational power to inductive bias ratio. Its wavy form maintains locality and introduces the new architectural streamlining, which allows it to be more adaptable to smartphone images. The observation is consistent with the previous results that convolutional inductive bias may be beneficial when tackling medical imaging problems that are characterized by noise and background variation.

Transformer based backbones have more mixed behavior. DeiT-Small and DINOv2-Small have comparatively small enhancements in the context of meta-learning and limited performance with the regular training, respectively, but significant improvements when combined with episodic learning. These trends indicate that task-level structuring of transformer representations might be needed to make them functional in very far out-of-domain applications.

5.5.3 Comparison Between Meta-Learning Paradigms

In this comparative analysis of meta-learning techniques, it becomes apparent that the principal benefits of meta-learning of episodic training as opposed to the particular meta-learning head selection. The performance of unseen-domain Prototypical Networks is always enhanced across all backbones, and this fact suggests that metric-based inference using episodic sampling forms a formidable baseline of zero-shot generalization.

Other more expressive meta-learning techniques, like Meta-Baseline, FEAT, and MetaOptNet, provide incremental advantages in some instances, but fail in general to compete with Prototypical Networks. Meta-Baseline outperforms by a significant margin other methods in terms of unseen performance on ConvNeXt-Tiny, although there are smaller or less reliable improvements by other methods.

These results indicate that in extreme cross-domain settings, less sophisticated meta-learning methods can be more consistent than highly parameterized or optimization-intensive methods. The observation is in line with the previous works who have found diminishing returns to the increase in meta-model complexity when the underlying domain change is large.

5.5.4 Comparison With Prior Work

Compared to existing literature on skin lesion classification, this study adopts a more stringent evaluation setting. Many prior works report high performance on dermoscopic datasets using either supervised learning or few-shot adaptation, often relying on partial access to target-domain data or closely matched acquisition conditions.

In contrast, the proposed evaluation framework enforces a strict zero-shot protocol in which no target-domain samples–labeled or unlabeled–are available during training or model selection. Under this constraint, absolute performance values are expected to be lower. However, the relative improvements observed through meta-learning are more meaningful, as they reflect genuine cross-device generalization rather than adaptation.

The results of this study therefore complement existing work by highlighting the limitations of source-optimized training and demonstrating the potential of episodic meta-learning to improve robustness in realistic deployment scenarios.

5.5.5 Summary of Observed Relationships

In summary, several key relationships emerge from the experimental analysis:

- High source-domain performance does not guarantee robust target-domain generalization.
- Meta-learning reduces the generalization gap between seen and unseen domains.
- Backbone architecture influences robustness, but increased capacity alone is insufficient.
- Simpler meta-learning methods often perform competitively under extreme domain shifts.
- Self-supervised representations benefit significantly from task-level structuring.

These relationships provide a coherent interpretation of the experimental results and reinforce the central conclusion that meta-learning offers a practical mechanism for improving zero-shot robustness in cross-device medical image classification.

5.6 Discussion

This part will summarize the experimental results of Chapter 5 and discusses the implications in the framework of zero-shot skin lesion classification when subjected to extreme cross-device domain shift. Instead of presenting new findings, the discussion is aimed at making sense of the trends observed, limiting it, and putting the contributions of this work in the perspective of realistic use.

5.6.1 Interpretation of Key Findings

The main conclusion to this research is that meta-learning can always enhance zero-shot generalization of unseen images learned in smartphones when trained against typical supervised training. In all considered backbone architectures, episodic meta-learning, especially the Prototypical Networks, lowers performance decline due to the dermoscopy-to-smartphone domain shift. Normal supervised models perform well with the source domain but do not perform similarly with the target domain. Such behavior does confirm that optimization on dermoscopic image statistics does not translate on the robustness of the image statistics when the acquisition conditions are not controlled. Conversely, meta-learning changes the optimization problem to be task-level generalization, which allows models to have more discriminative power when shifting domains. Notably, these gains are observed when no samples of the target-domain are available during training or evaluation. This makes the proposed methodology more practical as it represents the conditions of deployment where it is not possible to collect or annotate target-domain data.

5.6.2 Practical Implications for Cross-Device Deployment

In terms of deployment, the results indicate that there is an inherent problem with medical image analysis: models trained in a controlled clinical environment are not likely to work on consumer-grade imaging systems. It is hoped that smartphone images bring in diversity in illumination, background contents, scale, and noise that cannot be sufficiently represented by conventional dermoscopic datasets.

The reported gains obtained with the help of meta-learning indicate that episodic training can be a useful measure to address these problems. Although the performance on unseen devices is still not as high as in the source domain, the generalization gap is decreasing, which is a significant step to more reliable real-world systems.

These findings also suggest that enhancement of robustness to domain shift might be more critical than enhancing the performance of the source-domain accuracy, especially in the situations in which conditions of deployment are far apart in relation to the conditions of training.

5.6.3 Limitations of the Study

A number of weaknesses of this work should be supported. First, because of computational constraints, most experiments are run with a single random seed, and it is not possible to measure statistical variance between runs. Second, the analysis is confined to a set of backbone architectures and meta-learning techniques, and the findings might not be the same when applying other model designs.

Moreover, the rigid zero-shot condition means that the method of domain adaptation or test-time calibration which might further enhance the results cannot be used. Although this limitation makes the evaluation more valid, it hinders performance that can be attained.
Such limitations do not disqualify the findings of the study but they limit the scope through which the findings can be defined.

5.6.4 Contributions of This Work

This thesis has a number of contributions in its defined scope:

- It demonstrates that meta-learning consistently improves zero-shot generalization under severe dermoscopy-to-smartphone domain shift.
- It offers systematic testing of various backbone architectures in a rigid zero-shot protocol.
- It highlights the importance of episodic training over purely source-optimized supervised learning for cross-device robustness.
- It offers empirical insights into the interaction between backbone architecture and meta-learning strategy in medical image classification.

These contributions provide new insights to the current research by highlighting realistic deployment constraints as well as assessing robustness without target-domain adaptation.

5.6.5 Future Directions

This research indicates that the following lines of research can be taken in the future. Incorporating limited target-domain supervision or unlabeled data through few-shot adaptation or semi-supervised learning may further improve robustness. Exploring additional self-supervised pretraining strategies and backbone architectures could also yield more transferable representations.
Besides, it would be beneficial to test the proposed methodology on other medical imaging modalities and types of devices to determine their generality. Lastly, multi-seed analysis and more detailed statistical testing would be beneficial to the analysis to provide more confidence in the trends observed.

5.6.6 Concluding Remarks

Overall, it is possible to conclude that this thesis has shown that meta-learning is a viable and efficient methodology that can be used to enhance zero-shot robustness in cross-device skin lesion classification. Episodic learning is designed to halve the performance drop when models are applied to unknown devices by putting emphasis on generalization rather than source-domain optimization. Though the issue is still not resolved, the outcomes reported in this study are clear indications that meta-learning may be a viable path towards the security of addressing the gap between the controlled research standards and the actual use of the medical field.

Chapter 6

Conclusion

This thesis aimed to provide a narrow, deployment-oriented answer to the following question: can meta-learning be used to enhance strict zero-shot generalization when the models trained on dermoscopic images are used on clinical images taken with a smartphone? The experiments and analyses herein give a sophisticated and straightforward response: yes - episodic meta-learning reliably decreases the degradation brought about by dermoscopy-smartphone domain shift, although the quality of representation and experimental limitations dictate the magnitude of the gains. There are three themes that develop empirically. To begin with, there is the representational bottleneck. Standard supervised models that are trained on HAM10000 and BCN20000 source data show strong performance on source validation (macro-F1 0.70-0.83) and fail on the PAD-UFES-20 (macro-F1 0.20-0.29) indicating that high source accuracy does not follow through to cross-domain.This motivates the factorized experimental design: it is necessary but not sufficient to keep backbones fixed and simply switch the learning paradigm and observe that the quality of the representations can be predictive of cross-device transfer. Second, zero-shot robustness is systematically enhanced using episodic meta-learning. Across all tested backbones, prototypical episodic training increased unseen macro-F1 relative to supervised baselines (e.g., ConvNeXt-Tiny 0.29 - 0.35; ResNet 0.23 - 0.28), and the embeddings and inference processes created in this approach are more generalisable across modalities of acquisition. The benefit does not come solely as a result of a particular backbone: convolutional, transformer, and self-supervised structures were improved, and it would be reasonable to hypothesise that meta-learning decreases the use of source-specific information. Third, it does not provide much additional benefits making the meta-learning method more complex. The most significant enhancement of this is done by replacing normal supervised training with episodic meta-learning. Subsequently, more sophisticated meta methods like Meta-Baseline, FEAT or MetaOptNet merely generate minor improvements. In the case of ConvNeXt-Tiny, the difference between Meta-Baseline and ProtoNet (0.36 vs. 0.35 macro-F1 with unseen data) is quite small, but the benchmarking has significantly improved when switching to meta-learning. This demonstrates the fact that the way the model is trained on the task level is more significant than the complexity of the meta-learning head when dealing with a severe domain shift. Methodologically, the thesis emphasizes reproducibility and realism: a strict zero-shot protocol (no target images used for tuning or selection), a unified six-class label mapping, consistent preprocessing, and macro-F1 as the principal metric. Refinements Practical 6-way episodic tasks solved to avoid the exclusion of minority classes, group normalization to stabilize episodic training and robustness-oriented augmentations Materially enhanced stability and unobservable performance in repeated runs. These design decisions strengthen the causal claim that episodic objectives, not incidental tuning, drive the observed robustness gains. Limitations are acknowledged, computational restrictions constrained multi-seed evaluation and exhaustive tuning. The strict zero-shot set-up intentionally excludes domain adaptation techniques that might deliver higher absolute accuracy. The experiments cover a selected but representative set of backbones and meta-methods. These limitations imply that conclusions focus on similar trends as opposed to the supremacy of single-model. Meta-learning is a reliable mechanism to reduce the generalization gap, not eliminate it. There are two implications of the work in practice. To researchers: medical imaging system assessments should involve realistic cross-device tests; enhanced system performance on source benchmarks does not mean it is ready to be deployed. To practitioners and deployers: episodic meta-training can be a low-cost intervention (relative to collecting and labeling target data) to improve robustness when target supervision is unavailable. But it should be combined with careful backbone selection and lightweight robustness augmentations. Lastly, based on this thesis, a clear way forward in the future work is recognized: (1) incorporate limited target supervision (few-shot or semi-supervised) to quantify trade-offs between labeling cost and robustness; (2) evaluate larger pretraining regimes and multi-seed experiments for statistical confidence; (3) test on additional clinical modalities and devices to measure generality. The thesis not only provides valuable empirical data but also a practical paradigm of bringing medical image analysis to the point of safe and practical implementation by reformulating robustness as a collaborative inquiry of representation and learning goal. Episodic meta-learning proves to be the best in diminishing zero-shot degradation with a broad range of backbones; representation is the dominant lever; and careful experimental design under realistic constraints yields insights that are actionable for both research and applied deployment.

Bibliography

[1] A. Esteva, B. Kuprel, et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, pp. 115–118, 2017. DOI: 10.1038/nature21056 [Online]. Available: https://www.nature.com/articles/nature21056

[2] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International Conference on Machine Learning (ICML)*, 2017, pp. 1126–1135. [Online]. Available: https://arxiv.org/abs/1703.03400

[3] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4077–4087. [Online]. Available: https://arxiv.org/abs/1703.05175

[4] N. Codella et al., “Skin lesion analysis toward melanoma detection,” *Scientific Reports*, vol. 8, p. 13 458, 2018. DOI: 10.1038/s41598-018-30258-x [Online]. Available: https://www.nature.com/articles/s41598-018-30258-x

[5] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific Data*, vol. 5, p. 180 161, 2018. DOI: 10.1038/sdata.2018.161 [Online]. Available: https://arxiv.org/abs/1803.10417

[6] Y. Chen, Z. Liu, H. Xu, and T. Darrell, “A closer look at few-shot classification,” in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://arxiv.org/abs/1904.04232

[7] M. Combalia et al., “Bcn20000: Dermoscopic lesions in the clinical practice,” *arXiv preprint arXiv:1908.02288*, 2019. [Online]. Available: https://arxiv.org/abs/1908.02288

[8] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, “Metaoptnet: Differentiable convex optimization for few-shot learning,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 657–10 665. [Online]. Available: https://arxiv.org/abs/1904.03758

[9] Y. Guo, N. Codella, L. Karlinsky, et al., “A broader study of cross-domain few-shot learning,” *European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: https://arxiv.org/abs/1912.07200

[10] P. Khandelwal and P. A. Yushkevich, *Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging*, arXiv preprint arXiv:2008.07724, 2020. [Online]. Available: https://arxiv.org/abs/2008.07724

[11] W. Li, D. Xu, and H. Wang, "Few-shot learning as domain adaptation," *arXiv preprint arXiv:2002.02050*, 2020. [Online]. Available: https://arxiv.org/abs/2002.02050

[12] 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 (CVPRW)*, 2020, pp. 730–731. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w42/Mahajan_Meta-DermDiagnosis_Few-Shot_Skin_Disease_Identification_Using_Meta-Learning_CVPRW_2020_paper.html

[13] S. Mahajan, E. Tsironi, S. Gupta, and A. Chandrasekaran, "Meta-dermdiagnosis: Few-shot skin disease identification using meta-learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 0–0. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/papers/w42/Mahajan_Meta-DermDiagnosis_Few-Shot_Skin_Disease_Identification_Using_Meta-Learning_CVPRW_2020_paper.pdf

[14] A. G. C. Pacheco, B. Krohling, I. P. Biral, J. Pina, R. Ramos, and A. Rocha, "Pad-ufes-20: A skin lesion dataset composed of smartphone images," *Data in Brief*, vol. 32, p. 106 221, 2020. DOI: 10.1016/j.dib.2020.106221 [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7479321/

[15] E. Triantafillou, T. Zhu, V. Dumoulin, et al., "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://arxiv.org/abs/1903.03096

[16] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, vol. 53, no. 3, p. 63, 2020. [Online]. Available: https://arxiv.org/abs/1904.05046

[17] H. J. Ye, H. Hu, and D. C. Zhan, "Few-shot learning via embedding adaptation with transformer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 0–0. [Online]. Available: https://arxiv.org/abs/2001.03907

[18] R. R. Chowdhury and D. R. Bathula, *Influential prototypical networks for few shot learning: A dermatological case study*, arXiv preprint arXiv:2111.00698, 2021. [Online]. Available: https://arxiv.org/abs/2111.00698

[19] J. B. Grill, F. Strub, F. Altché, et al., "Self-distillation with no labels," in *International Conference on Computer Vision (ICCV)*, 2021. [Online]. Available: https://arxiv.org/abs/2104.14294

[20] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. [Online]. Available: https://arxiv.org/abs/2004.05439

[21] C. Magalhães, C. Eloy, A. Afonso, et al., “The role in teledermoscopy of an inexpensive and easy-to-use smartphone device for the classification of three types of skin lesions using convolutional neural networks,” *Diagnostics*, vol. 11, no. 3, p. 451, 2021. DOI: 10.3390/diagnostics11030451 [Online]. Available: https://doi.org/10.3390/diagnostics11030451

[22] R. Khadka et al., “Meta-learning with implicit gradients in a few-shot setting for medical image segmentation,” *Computers in Biology and Medicine*, vol. 143, p. 105 227, 2022. DOI: 10.1016/j.compbiomed.2022.105227 [Online]. Available: https://doi.org/10.1016/j.compbiomed.2022.105227

[23] J. Liu, Q. Wang, Z. Liu, et al., “A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone-collected clinical images and metadata,” *Frontiers in Surgery*, vol. 9, p. 1 029 991, 2022. DOI: 10.3389/fsurg.2022.1029991 [Online]. Available: https://doi.org/10.3389/fsurg.2022.1029991

[24] A. Yilmaz, G. Gencoglan, R. Varol, A. A. Demircali, M. Keshavarz, and H. Uvet, “Mobileskin: Classification of skin lesion images acquired using mobile phone-attached hand-held dermoscopes,” *Journal of Clinical Medicine*, vol. 11, no. 17, p. 5102, 2022. DOI: 10.3390/jcm11175102 [Online]. Available: https://doi.org/10.3390/jcm11175102

[25] K. Lee et al., “Multi-task and few-shot learning-based fully automatic deep learning platform for mobile diagnosis of skin diseases,” *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 176–187, 2023. DOI: 10.1109/JBHI.2022.3193685 [Online]. Available: https://doi.org/10.1109/JBHI.2022.3193685

[26] J. Nayem et al., “Few shot learning for medical imaging: A comparative analysis of methodologies and formal mathematical framework,” in *Data Driven Approaches on Medical Imaging*, Springer, 2023, pp. 69–90. DOI: 10.1007/978-3-031-47772-0_4 [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-47772-0_4

[27] M. Oquab, T. Darcet, et al., “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023. [Online]. Available: https://arxiv.org/abs/2304.07193

[28] S. Pati et al., “A systematic review of few-shot learning in medical imaging,” *arXiv preprint arXiv:2309.11433*, 2023. [Online]. Available: https://arxiv.org/abs/2309.11433

[29] M. Azeem, K. Kiani, T. Mansouri, and N. Topping, “Skinlesnet: Classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional neural network,” *Cancers*, vol. 16, no. 1, p. 108, 2024. DOI: 10.3390/cancers16010108 [Online]. Available: https://doi.org/10.3390/cancers16010108

[30] S. Chamarthi, K. Fogelberg, J. Gawlikowski, and T. J. Brinker, “Few-shot learning for skin lesion classification: A prototypical networks approach,” *Informatics in Medicine Unlocked*, vol. 48, p. 101 520, 2024. DOI: 10.1016/j.imu.2024.101520 [Online]. Available: https://doi.org/10.1016/j.imu.2024.101520

[31] S. Chamarthi, K. Fogelberg, R. C. Maron, T. J. Brinker, and J. Niebling, "Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods on dermoscopic images," *Informatics in Medicine Unlocked*, vol. 44, p. 101 430, 2024, Also available as arXiv:2310.03432. DOI: 10 . 1016 / j . imu . 2023 . 101430 [Online]. Available: https://doi.org/10.1016/j.imu.2023.101430

[32] T. Chen, Y. Tian, Y. Chen, et al., "Few-shot classification with multiscale feature fusion for clinical skin disease diagnosis," *Clinical, Cosmetic and Investigational Dermatology*, vol. 17, 2024. DOI: 10.2147/CCID.S458255 [Online]. Available: https://doi.org/10.2147/CCID.S458255

[33] M. García, S. A. Kostopoulos, C. Mariño, et al., "Optimizing digital image quality for improved skin cancer detection," *Journal of Imaging*, vol. 11, no. 4, p. 107, 2025. DOI: 10.3390/jimaging11040107 [Online]. Available: https://doi.org/10.3390/jimaging11040107

[34] T. Kränke, H. Willschke, T. Tzellos, et al., "Assessment of a smartphone-based neural network application for the risk assessment of skin lesions under real-world conditions," *Dermatology Practical & Conceptual*, vol. 15, no. 4, e2025a198, 2025. DOI: 10.5826/dpc.1504a198 [Online]. Available: https://doi.org/10.5826/dpc.1504a198

[35] S. Yan, X. Li, M. Hu, Y. Jiang, Z. Yu, and Z. Ge, *Make: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment*, arXiv preprint arXiv:2505.09372, 2025. [Online]. Available: https://arxiv.org/abs/2505.09372