Informatics in Medicine Unlocked 48 (2024) 101520

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Few-shot learning for skin lesion classification: A prototypical networks approach

Sireesha Chamarthi [a,1], Katharina Fogelberg [b,1], Jakob Gawlikowski [a], Titus J. Brinker [b,*]

[a] German Aerospace Center (DLR), Institute of Data Science, Jena, Germany
[b] Digital Biomarkers for Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany

ARTICLE INFO

Keywords:
Few-shot learning
Domain shift
Skin lesion classification
Dermoscopic images
Domain generalization

ABSTRACT

Prototypical networks (PN) have emerged as one of multiple effective approaches for few-shot learning (FSL), even in medical image classification. This study focuses on implementing a PN for skin lesion classification to assess its performance, generalizability, and robustness when applied across 11 dermoscopic image domains. Unlike conventional FSL scenarios, where the performance is evaluated for unseen classes in the test set, our analysis extends this to evaluate PNs on a complete hold-out dataset with the same classes from a different domain. Differences in a patient's age, lesion localization, or image acquisition systems variations mimic real-world cross-domain conditions in a clinic. Given the scarcity of medical datasets, this assessment is crucial for potentially translating such systems into real-world clinical settings to support physicians with the diagnosis. Our primary focus is two-fold: investigating whether a PN performs on par with a baseline classifier, even using only a limited number of reference samples from the hold-out test set (in-domain) and whether a PN can generalize to the same classes of unseen domains (cross-domain). Our analysis uncovers that a PN can perform on par with the baseline classifier in an in-domain setting, even with only a few support samples. However, in cross-domain scenarios, a PN exhibits improved performance only on specific domains, while others demonstrate similar or even decreased performance when confronted with a smaller number of images. Our findings contribute to comprehending potential opportunities and limitations of FSL in dermatological practice.

1. Introduction

Deep Learning (DL) techniques have been widely used in diverse medical imaging tasks, including classification and segmentation [1]. In particular, in skin lesion classification, deep learning models have shown promising results [2]. However, two significant obstacles are preventing the utilization of these techniques in clinical practice: the scarcity of medical data, particularly labeled and rare data, and challenges related to the generalization across different domains with a domain shift present. Domain shifts arise when the training dataset of the classification model is from a different distribution than the testing dataset. This is a typical scenario in clinical skin cancer diagnosis due to the differences in image acquisition systems or different patient groups, as shown in our earlier analysis [3].

Several methods deal with data scarcity in machine learning, e.g., data augmentation, transfer learning, and few-shot learning (FSL). The latest research describes FSL as a promising approach for medical image classification [4] because it focuses on using only the image features of a few samples (shots) per class for model training.

To address the limitation of handling domain shifts, transfer learning approaches like domain adaptation or domain generalization can be employed [5]. Domain adaptation is used when the task of the model remains the same while the distribution between two datasets (source- and target domain) differs [6]. In contrast, domain generalization uses a model to train on multiple source domains with different data distributions, thus improving the generalization capabilities of the model when applied to an unseen dataset (target domain) [5].

The success of transfer learning approaches depends on the availability of data in two (or more when using domain generalization) domains during the training of the models. If fine-tuning as a transfer learning approach is used, the models tend to overfit on small datasets [7]. When domain adaptation is used to address domain shifts, success relies on large amounts of training data from the source and

[*] Correspondence to: Division of Digital Biomarkers for Oncology, GermanCancer Research Center (DKFZ), Heidelberg, Germany.
E-mail address: titus.brinker@nct-heidelberg.de (T.J. Brinker).

[1] Both authors contributed equally.

https://doi.org/10.1016/j.imu.2024.101520
Received 21 November 2023; Received in revised form 8 May 2024; Accepted 9 May 2024
Available online 11 May 2024

target domain, thus leading to a long duration of adaptation. This highlights the need for techniques that can handle limited data and adaptation to new domains. While FSL has primarily been employed in in-domain scenarios to address data scarcity [8–10], we aim to examine its potential in cross-domain settings to tackle data scarcity and generalization. For this task, we employed a Prototypical Network (PN) due to its broad usage in the medical field [11–15].

Our contribution lies in investigating the generalization capabilities of PNs on cross-domain dermoscopic images, covering the spectrum of shifts across clinic- and patient-specific scenarios. This explores the adaptability of PNs across cross-domain images. For this purpose, we initially assessed the performance of PNs against a baseline classifier specifically tailored to in-domain skin lesion datasets. Additionally, we evaluated the impact of various hyperparameters (episodes, epochs, shots, and training layers) on the performance of the PN, particularly within the context of in-domain and cross-domain dermoscopic images. This exploration goes beyond the baseline model, offering insights into the dynamics that influence the model's performance.

Section 2 discusses relevant research in meta-learning, FSL, and using PNs in dermoscopic scenarios. Subsequently, in Section 3, we provide details about the datasets and models we utilized, including a baseline classifier and a PN. In this regard, we explain the meta-training and cross-domain meta-testing processes. Within Section 4, we present an evaluation of PNs, encompassing their performance within in-domain and cross-domain settings, along with the influence of diverse hyperparameters on PN performance. Moreover, a comparative analysis between the FSL model and a baseline classifier is presented. We summarize this study's main findings in Section 5.

2. Related work

Most FSL methods belong to the branch of meta-learning, known as "learning to learn" [16], which involves teaching a model multiple tasks to improve its ability to quickly adapt to an entirely new task [17–19]. Parnami & Lee classified few-shot meta-learning approaches into metric-, optimization- and model-based methods, depending on how the learning task is defined [19].

Optimization- or gradient-based approaches perform by implementing changes to the network optimization process, of which Model-Agnostic Meta-Learning (MAML) [20] is a popular method. With the model-based approach Simple Neural Attentive Learner (SNAIL) [21], learning is achieved by combining experience aggregation and attention. Metric-based approaches are popular for measuring the similarity or distance between samples, aiming to create a metric space where samples from the same class are brought closer and samples from a different class are far apart. The most prominent methods for metric learning are Siamese networks [22], Matching networks [23], Relation networks [24], and Prototypical networks [8].

Significant progress has been made in using meta-learning to adapt unseen domains. One approach focuses on supervised domain adaptation when only a limited amount of labeled target samples is accessible, making it applicable to FSL scenarios [25]. An alternative strategy was presented by Sahoo et al. proposing a combination of meta-learning and strategies to mitigate domain shift with adversarial domain adaptation [26]. Also, the customization of a PN by fine-tuning its backbone was presented as a valid approach for domain adaptation with FSL [27]. This work additionally conducted ablation studies focusing on different hyperparameters of PNs. Laenen et al. also demonstrated that different hyperparameters affect the performance of a PN [28].

Recent studies indicate the successful adoption of FSL methods in skin lesion classification. Liu et al. used an improved version of Relation Networks for skin disease classification [29]. Also, a gradient-based meta-learning approach has been proposed for the classification of medical images [30]. Furthermore, PNs gained popularity in dermatological diagnosis. For instance, Mahajan et al. proposed a method called Meta-Derm-Diagnosis on skin lesion datasets with limited annotated examples, which is employed with Reptile and PNs [11]. Furthermore, Prabhu et al. used FSL for dermatological disease diagnosis by introducing Prototypical Clustering Networks based on PNs [15]. In their case, skin lesions are classified by a similarity measure of weighted combinations of prototypes for a class. While existing research has shown success in few-shot learning for dermatological cases, our approach aims to push the boundaries by extending FSL experiments to cross-domain scenarios. This leads us to our fundamental question: Can FSL models effectively generalize to cross-domain dermatological applications?

3. Materials and methods

3.1. Prototypical networks for few shot learning

In Prototypical Networks (PNs), the meta-learning process involves two key steps: meta-training and meta-testing [8,16]. Fig. 1 shows the schematic of the meta-learning. An episode in FSL is a single learning task that consists of a support set and a query set. The support set contains a small number of samples (shots) from different classes that the model uses to learn the task. An epoch in FSL is typically represented by a complete iteration over the entire set of available episodes. The meta-training phase aims to train a model that can adapt to new tasks or domains with only a few examples (few-shot learning). Like meta-training, each meta-testing episode involves a support set and a query set. However, during meta-testing, the model is presented with tasks not seen during the meta-training phase.

As shown in Fig. 2 in the meta-training phase, support and query set images are projected by a feature extractor into an embedding space. This projection is performed for each episode and is characterized by a small number of support set images. An episode has five samples from each class (Nevus and Melanoma). The model calculates the prototype for each class by computing the mean of the support set samples. Prototype of support samples for each class P_s:

$$P_s = \frac{1}{|S_c|} \sum_{(x_i,y_i)\in S_c} f_\theta(x_i) \tag{1}$$

where S_c is the support samples for class c and f_θ is the embedding of all the support samples. Later, the projected query sample is assigned to the class of the closest prototype based on a distance metric. The distance metric we employed is based on the Euclidean measure. The model incorporates cross-entropy loss to categorize and generalize in few-shot learning situations.

Our approach is to train the model(s) exclusively on data from a single domain during the meta-training stage. In the meta-testing phase, the model's performance is evaluated on new patient groups and previously unseen clinical settings. This evaluation mirrors conventional transfer learning, where a model is initially trained on one domain (source) and then fine-tuned on another (target). However, in this study, we aim to determine the feasibility of transferring knowledge and adapting to an entirely unfamiliar domain (target) using only a limited number of images.

3.2. Datasets

In our prior study [3], we categorized[2] three large ISIC datasets: HAM [31], BCN [32], and MSK [33] as technical (clinic-specific) and biological (patient-specific) domains. Thus, the three datasets are further divided into sub-datasets based on their domain shifts. Domain shifts in these datasets arise primarily due to (a) the variations arising from changes in the origin of the dataset (different clinics with different image acquisition systems) and (b) the other category arises from differences in age and location of the skin lesions of patients. Table 1 shows

[2] https://gitlab.com/dlr-dw/isic_download

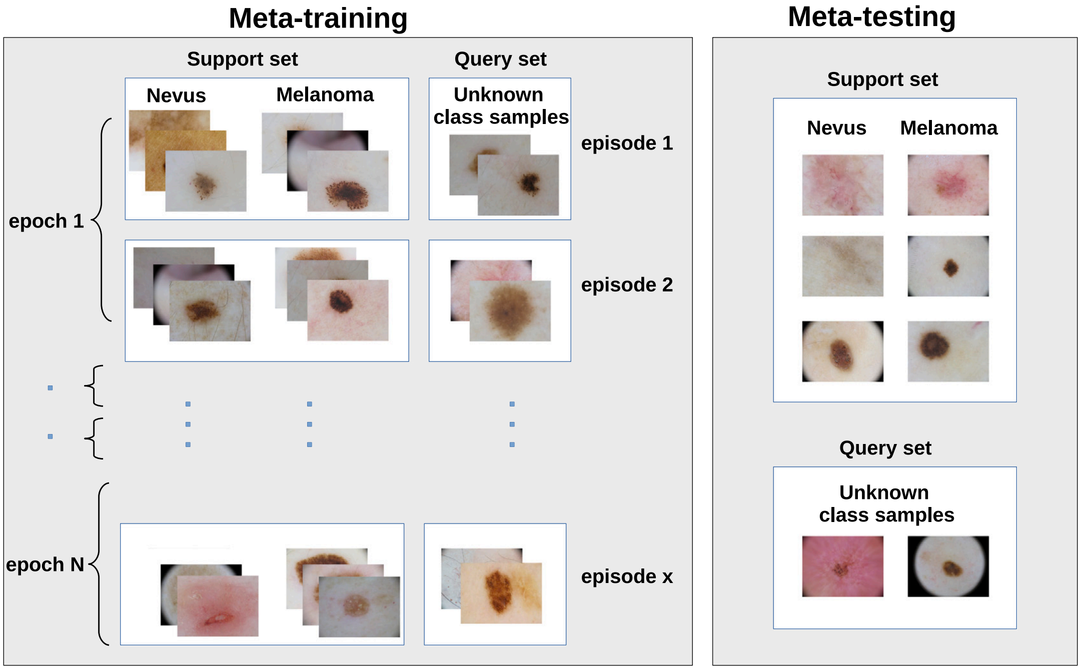

Fig. 1. Schematic to demonstrate the meta-learning approach. The meta-training approach comprises several episodes per epoch. Each episode consists of support and query sets. Each support set comprises a few samples (shots) from each class. The class label for the query set is assigned based on the closest prototype of support set samples. In the meta-testing phase, unseen data from in- or cross-domain sources is used.

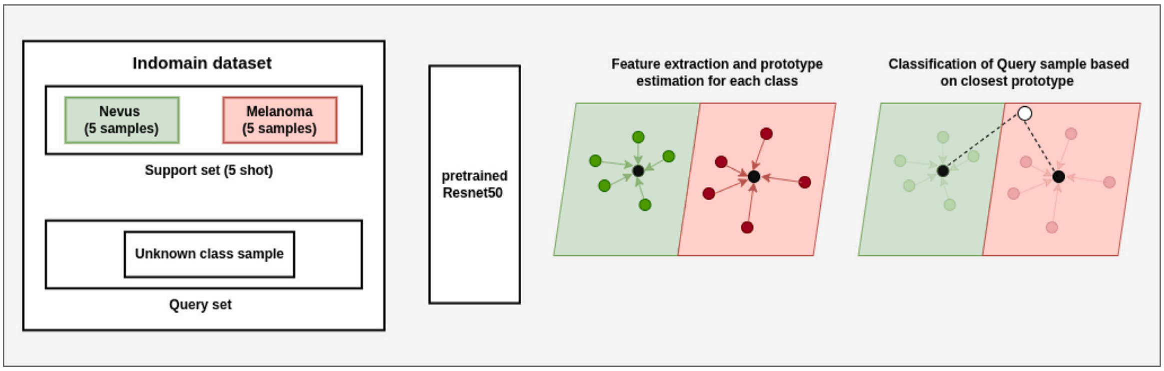

Fig. 2. Schematic demonstrating the approach of Prototypical networks. The samples from the support set are projected using a pre-trained ResNet50 model. The prototype is computed for each class after feature extraction. The query sample is assigned a class label based on the closest prototype to the query projection.

the domain characteristics we used to divide the three datasets into different domains. The most common clinic domain is represented using groups with patients *aged over 30 with the lesion localization on the torso*. Other domains include *the same age group but with lesion localization at the head/neck* and *palms/soles*. *Age less than 30* is considered as a different domain. Each domain is further split into support and query datasets with an equal distribution between both classes, as shown in Table 1.

3.3. Meta-training

With the episodal training, the model was trained to predict query labels based on support images from each class in the domain. Our training domain comprises the group with *age over 30 with lesion localization on the torso*. We employed a ResNet50, pre-trained on ImageNet without fine-tuning, to project support and query images into the embedding space.

Due to the balanced training sets by class, we calculated the average accuracy over multiple episodes. For comparison, we also calculated the average AUROC, which is threshold-free and widely used in medical settings. However, as our datasets are balanced, (a) accuracy is the most feasible measure, and (b) the results from AUROC were matching with the accuracy output. As outlined in [28], this episodal training involves a combination of hyperparameters such as episodes, shots, etc. Following established practices [18], we assessed performance using episodes for 2-way classification (melanoma and nevus) and conducted experiments with 10-, 5-, 3-, and 1-shot scenarios.

3.4. Clinic- and patient-specific meta-testing

We evaluated the PNs' performance in two scenarios: in-domain and cross-domain. Throughout these evaluations, we maintained a consistent number of shots (support samples) for both training and testing.

For in-domain meta-testing, we assessed the model's performance on unseen hold-out data from the same domain and the same dataset as that of training, as shown in Table 2. For this, we partitioned each domain of the original datasets (HAM, BCN, and MSK) into meta-training (Train-domain) and meta-testing (In-domain) in an 80:20 ratio as shown in Table 2.

Table 1
Overview of the datasets used for training and testing. The datasets HAM, BCN, and MSK are further partitioned into distinct domains based on specified features in the *Domain characteristics* column. The corresponding support and query set sizes are shown in the last column. The support- and query images are further partitioned into small episodes while maintaining a balanced class ratio.

Dataset origin	Domain characteristics	Support/Query	Dataset size
HAM	age > 30, loc. = body (default)	Support set	745
		Query set	185
	age ≤ 30, loc. = body	Support set	41
		Query set	9
	age > 30, loc. = head/neck	Support set	159
		Query set	39
	age > 30, loc. = palms/soles	Support set	25
		Query set	5
BCN	age > 30, loc. = body (default)	Support set	3070
		Query set	765
	age ≤ 30, loc. = body	Support set	114
		Query set	28
	age > 30, loc. = head/neck	Support set	513
		Query set	127
	age > 30, loc. = palms/soles	Support set	169
		Query set	41
MSK	age > 30, loc. = body (default)	Support set	905
		Query set	225
	age ≤ 30, loc. = body	Support set	61
		Query set	13
	age > 30, loc. = head/neck	Support set	188
		Query set	46

In cross-domain testing, we considered *clinic-specific* and *patient-specific* aspects. For *clinic-specific* cross-domain experiments, the corresponding test dataset is sourced from the same domain but a different dataset (for HAM, BCN, MSK). Table 2 shows the meta-training (Train-domain) and meta-testing (Cross-domain) datasets used in *clinic-specific* experiments. In this scenario, an example is training on BCN and subsequently testing on HAM from the same domain. *Clinic-specific* meta-testing consisted of 500 episodes with five shots for support- and query images.

Table 3 shows the *patient-specific* experiments, where the differentiation lies in utilizing distinct domains from the same dataset for the training and testing phases. An example is to train on BCN and test on one of the BCN patient-domain datasets. Patient-specific domains, characterized by smaller melanoma and nevus distributions, underwent testing with 500 episodes and two shots.

While hyper-parameter variations were explored for meta-testing scenarios, they had minimal impact on the performance, leading us to adhere to the mentioned settings. The model performance for in-domain and cross-domain scenarios was assessed using average accuracy and AUROC across all episodes. To evaluate the repeatability and uncertainty in the results, we calculated the accuracy over three seeds with the corresponding mean and standard deviation of the results.

3.5. Baseline classifier

We additionally evaluated the performance of a PN in comparison to a baseline classifier. For this purpose, we utilized a ResNet50 [34] model pre-trained on the ImageNet dataset as our backbone on each domain listed in Table 1. ResNet50 is well established for its effectiveness as a feature extractor, which is crucial for extracting the intricate patterns and features within skin lesion images. Leveraging a pre-trained model facilitated knowledge transfer from a diverse range of images, enhancing the network's capacity to learn and generalize across different skin lesion domains. This choice of architecture aligns with the objective of achieving a robust classifier for skin lesion analysis. We used the same data for PNs and the baseline classifier to maintain the comparison fair. In the case of the FSL model, the model learns to classify the unseen sample even with a few shots of data (1, 3, 5, and 10). The model uses the entire support set for training and the query set for validation for the baseline classifier. For instance, if HAM (loc = body, age >30) serves as the training domain and BCN (loc = body, age >30) is the test domain. The support set for the meta-training of the FSL model is utilized as the training set for the baseline classifier. The corresponding query set for the FSL model is employed as the validation set for the baseline model. Finally, the query set in meta-testing for the FSL model is used as the test set for the baseline classifier.

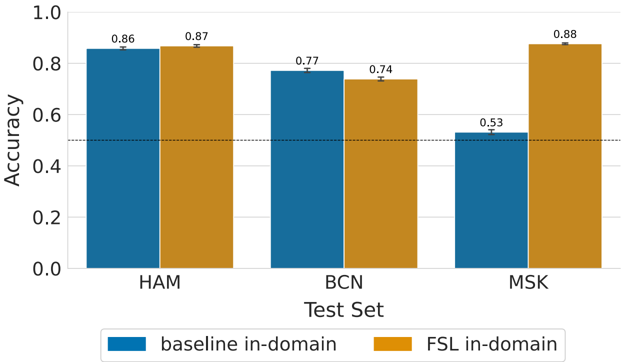

Fig. 3. Performance comparison of the baseline and the FSL method in an in-domain setting. The in-domain dataset used in this experiment is shown in Table 2. This plot shows that FSL shows similar/improvement in performance for most of the in-domain datasets.

4. Results and discussion

In the following sections, we present the results of our analysis, where we explored PNs and their ability to perform on in-domain and cross-domain dermoscopic images compared to a baseline classifier (ResNet50). This study aims to assess the PNs' adaptability to datasets within unseen clinic and patient-specific domains. Our analysis involved a series of experiments designed to observe how different hyper-parameters influence the performance of a PN. We then compared the performance and generalizability of this optimized PN configuration against a baseline classifier (ResNet50).

It is important to note that our primary focus in this analysis is not the performance of the baseline classifier or enhancing the performance of the baseline classifier. Several works have already established the efficacy of a baseline classifier with many datasets and its performance on an unseen dataset from the same domain [2]. We want to simulate a real-world scenario where obtaining a large dataset for training is often impractical and expensive. Hence, we used the same small dataset to train the FSL models. Also, this would ensure a fair comparison with the FSL model performance, avoiding any potential bias from a substantial disparity in the training data. The baseline classifier serves as a benchmark for comparison, assessing how a pre-trained ResNet50 model functions when provided with the same dataset.

4.1. In-domain performance

Dermoscopic image classifiers achieve good results when tested in ideal experimental settings, such as data from the same distribution. We can also observe this pattern in our results in Fig. 3. Typically, a large amount of data is used to train such classifiers, unlike FSL methods, which only use a few images. As shown in Fig. 3, the performance of the PN and the baseline classifier appear similar in two out of three cases. However, the baseline classifier seems to perform poorly for the MSK dataset, whereas the FSL model shows substantial performance

Table 2
Clinic domains used in the experiments. From each of the training domains, 20% of the hold-out dataset is used for in-domain tests for both support and query sets. Whereas for cross-domain testing, the same domain from different data origins (clinic-specific) is used for evaluation.

Train data	Test data	
Train domain	In-domain	Cross-domain
HAM (loc=body, age > 30)	HAM (loc=body, age > 30)	MSK (loc=body, age > 30) BCN (loc=body, age > 30)
MSK (loc=body, age > 30)	MSK (loc=body, age > 30)	HAM (loc=body, age > 30) BCN (loc=body, age > 30)
BCN (loc=body, age > 30)	BCN (loc=body, age > 30)	HAM (loc=body, age > 30) MSK (loc=body, age > 30)

Table 3
Patient domains used in the experiments. Patient cross-domains are selected within the same dataset group but for a different domain.

Train data	Test data (Cross-domain)
HAM (loc=body, age> 30)	HAM (age ≤ 30, loc. = body) HAM (age > 30, loc. = head/neck) HAM (age > 30, loc. = palms/soles)
BCN (loc=body, age > 30)	BCN (age ≤ 30, loc. = body) BCN (age > 30, loc. = head/neck) BCN (age > 30, loc. = palms/soles)
MSK (loc=body, age > 30)	MSK (age ≤ 30, loc. = body) MSK (age > 30, loc. = head/neck)

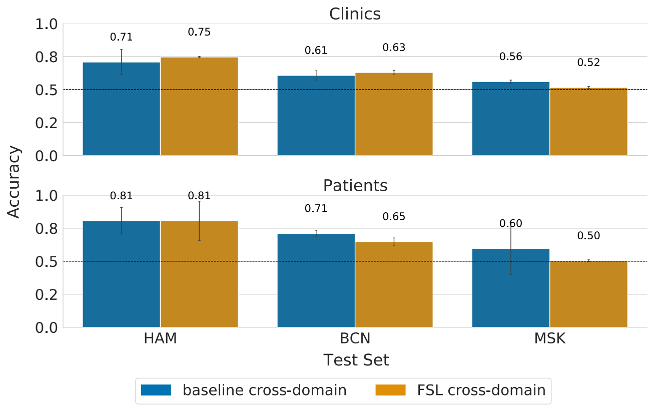

Fig. 4. Performance comparison of the baseline and the FSL method in a cross-domain setting. The top plot compares the baseline classifier with FSL for clinic-cross domains shown in Table 2. The bottom plot shows the comparison for patient-cross domains shown in Table 3.

improvement. This illustrates that a limited amount of images with a PN is sufficient to achieve comparable performance to that of the skin cancer classification's baseline classifier. These findings indicate a promising direction for the potential applicability of FSL in medical practice, particularly in scenarios where data availability is limited.

4.2. Cross-domain performance

In potential real-world scenarios, the setting is often less than ideal. Frequently, the distribution between the training and test set differs, as classifiers are evaluated on novel and diverse cases they have not encountered before. Consequently, the assumption is that the performance decreases when testing the same classifier on a cross-domain dataset. We can observe this decrease in performance for the baseline classifier and the PN, decreasing from in-domain Fig. 3 to cross-domain Fig. 4. As we are particularly interested in the cross-domain performance of the FSL model, we observed the behavior on clinic- and patient-specific domain shifts separately.

Even in cross-domain scenarios, the performances of the baseline classifier and the FSL model appear similar (Fig. 4). In general, achieving high accuracy in patient domains indicates a more effective adaptation to domain shifts for both the baseline classifier and Prototypical Networks (PN), in contrast to shifts observed between different clinics. This effectiveness is likely due to the smaller domain shift arising from biological differences, which proves more manageable than the technical differences encountered [3].

Deciding whether PNs can be easily employed to adapt to a new domain is challenging because no clear pattern can be identified from the results. Although the performance is generally superior for patient domain shifts, for BCN and MSK, the cross-domain performance decreases compared to the baseline classifier. For HAM, the performance remains unchanged. Conversely, when analyzing clinical domain shifts, the performance increases for HAM and BCN when a PN is used. Only for the generally challenging dataset MSK a decrease in performance can be observed with a PN. Overall, employing a PN seems to yield more stable results, as the standard deviation is consistently lower.

4.3. Effects of different hyperparameters

According to Laenen et al. [28], the different parameters that are involved in episodic training, specifically ways (classes), episodes (support-query pairs), and shots (number of support images for training), can significantly affect performance. Consequently, a comprehensive understanding of the impact these hyperparameters and others have on the performance is essential. Therefore, we experimented with varying numbers of episodes and shots. However, we did not explore the impact of different class quantities on performance, as our focus remains on the binary classification of melanomas and nevi in dermoscopic images to facilitate comparison. Instead, we additionally investigated epochs and the usage of different network layers during training, recognizing their typically important role in classification tasks.

In Fig. 5, it is evident that, overall, the distinctions between various hyperparameter values are not substantial. Specifically, the differences in average cross-domain accuracy prove to be negligible. For in-domain scenarios, the accuracy shows minor fluctuations depending on the hyperparameter values, except for the epochs, where it remains relatively constant. This indicates no considerable correlation between the hyperparameters and performance, a conclusion that we can confirm through correlation matrices. However, the in-domain accuracy seems weakly influenced (r = 0.35) by the number of episodes used, which cannot be observed for the cross-domain setup. Moreover, the accuracy does not exhibit significant variation across the three runs, with only a slightly larger standard deviation observed when using different layers on the BCN test sets.

As shown in Fig. 5 (top left), employing a modest episode count of only 100 yields a notable accuracy and reaches a saturation point from 500 episodes onward. In this context, despite the potential use of more episodes in other studies, the limited size of our datasets suggests that utilizing only 500 episodes is sufficient. As illustrated in Fig. 5 (top right), the amount of shots exhibits minimal variation.

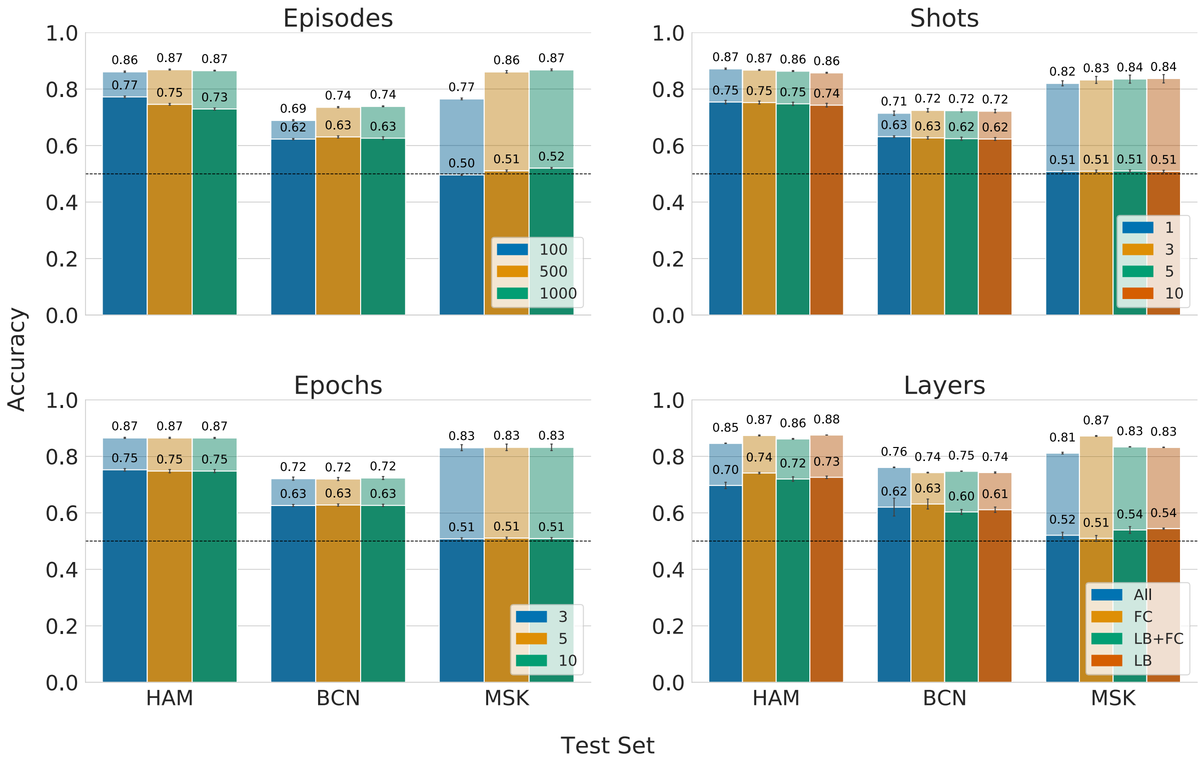

Fig. 5. Effect of different hyperparameters on the performance of a PN tested on clinical cross-domain datasets. Transparent bars represent in-domain (higher) performance, while the opaque bars in the front represent cross-domain (lower) performance. Confidence intervals across three runs.

It is noteworthy that in other FSL studies, various shot quantities are often compared for a single task to assess how their performance differs. Conventionally, the expectation is that using only one shot for a task would result in considerably lower accuracy. However, our analysis reveals no substantial correlation between shots and accuracy, especially when considering cross-domain scenarios. In Fig. 5 (bottom left), there is no noticeable difference in performance when varying the epoch size. As known in the machine learning community, selectively training specific blocks or layers of neural networks can impact performance and significantly decrease computation time. In our experiments, we found that training all layers does not yield optimal results. Instead, the effectiveness depends on the test datasets. Specifically, for HAM and MSK in cross-domain scenarios, training only the fully connected layer while freezing the weights of the other network layers produces the best results, as illustrated in Fig. 5 (bottom right). Additionally, our findings on distance metrics used for PNs align with those of Snell et al. [8], showing superior performance with Euclidean distance than with cosine distance in our classification scenario.

In summary, the results collectively indicate that classifying cross-domain MSK data is generally more challenging, while HAM data achieves the most favorable performance results. By averaging results across three runs, Fig. 5 shows a slight decrease in performance when transitioning from the in-domain to the cross-domain scenarios for the HAM dataset, with the most notable decrease observed for MSK data. Interestingly, the number of epochs does not appear to be important because the model has already learned effectively by the third epoch and only has 500 episodes. Consequently, PNs can learn from relatively small datasets, performing comparably to the baseline classifier.

From Figs. 3 and 4, particularly for the MSK dataset, the results suggest that PN shows improvement in in-domain performance even with small datasets. However, when tested on cross-domain, the performance is sub-optimal. Our earlier analysis showed that MSK is a difficult to adapt domain even while training with the full dataset [3]. These results are in accordance with what was observed earlier. It is worth mentioning that one of the subdomains within MSK (age >30, loc=head/neck) proved to be particularly challenging for several unsupervised domain adaptation methods [35]. Further investigation is required to understand how to effectively adapt the MSK dataset, a task that extends beyond the scope of this manuscript.

While this represents the initial step in assessing the performance of PNs across domains, we acknowledge the importance of addressing these complexities and recognize the need for future research efforts in this direction. One potential avenue for exploration could involve multimodal few-shot learning, which enhances adaptation strategies and improves performance in challenging domains such as MSK. Another interesting approach for such datasets would be to utilize weighted prototypical networks, emphasizing intra-class distribution [36]. Additionally, conducting a statistically significant analysis of the employed datasets is essential to implement real-time diagnostic assistant systems.

5. Conclusion

Throughout this research, we have uncovered several important findings on the performance, adaptability, and robustness of PNs across multiple domains. The similarity between the baseline- and FSL performance suggests that FSL could be applicable in practice within the same domain and in rare cross-domain cases. While analyzing our data, we observed that epochs, episodes, and shots do not correlate strongly with accuracy, indicating that they do not considerably impact performance. However, episodes and freezing of different parts of the network layers can weakly influence performance. Additionally, there is no clear indication that clinic- or patient-specific domain shifts are easier to adapt to. Nevertheless, when evaluating cross-domain performance, we observe that both the baseline- and FSL models exhibit slightly better performance when applied to patient-specific domain shifts. The results of this study can be used as a direction to the potential limitations and opportunities in clinical decision-making, especially with limited data.

CRediT authorship contribution statement

Sireesha Chamarthi: Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Katharina Fogelberg:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Jakob Gawlikowski:** Validation, Supervision, Resources, Project administration, Investigation, Formal analysis. **Titus J. Brinker:** Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Titus Josef Brinker would like to disclose that he is the owner of Smart Health Heidelberg GmbH (Handschuhsheimer Landstr. 9/1, 69120 Heidelberg, Germany) which develops mobile apps, outside of the submitted work.

Acknowledgments

This research is funded by the *Helmholtz Artificial Intelligence Cooperation Unit* [grant number ZT-I-PF-5-066].

The Helmholtz AI funding enabled the close cooperation between DKFZ and DLR, which leads to an interdisciplinary exchange between two research groups and thereby enables the integration of novel perspectives and experiences.

References

[1] Razzak Muhammad Imran, Naz Saeeda, Zaib Ahmad. Deep learning for medical image processing: Overview, challenges and the future. In: Lecture notes in computational vision and biomechanics. Springer International Publishing; 2017, p. 323–50. http://dx.doi.org/10.1007/978-3-319-65981-7_12, Retrieved on April 19, 2024.

[2] Esteva Andre, Kuprel Brett, Novoa Roberto A, Ko Justin, Swetter Susan M, Blau Helen M, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8. http://dx.doi.org/10.1038/nature21056, Retrieved on April 19, 2024.

[3] Fogelberg Katharina, Chamarthi Sireesha, Maron Roman C, Niebling Julia, Brinker Titus J. Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. New Biotechnol 2023;76:106–17. http://dx.doi.org/10.1016/j.nbt.2023.04.006, Retrieved on April 19, 2024.

[4] Nayem Jannatul, Sahriar Hasan Sayed, Amina Noshin, Das Bristy, Shahin Ali Md, Manjurul Ahsan Md, et al. Few shot learning for medical imaging: A comparative analysis of methodologies and formal mathematical framework. 2023, http://dx.doi.org/10.48550/arXiv.2305.04401, arXiv e-prints, arXiv:2305.04401. Retrieved on April 19, 2024.

[5] Wang Jindong, Lan Cuiling, Liu Chang, Ouyang Yidong, Qin Tao. Generalizing to unseen domains: A survey on domain generalization. In: Proceedings of the thirtieth international joint conference on artificial intelligence. International Joint Conferences on Artificial Intelligence Organization; 2021, http://dx.doi.org/10.24963/ijcai.2021/628, Retrieved on April 19, 2024.

[6] Wang Mei, Deng Weihong. Deep visual domain adaptation: A survey. Neurocomputing 2018;312:135–53. http://dx.doi.org/10.1016/j.neucom.2018.05.083, Retrieved on April 19, 2024.

[7] Yosinski Jason, Clune Jeff, Bengio Yoshua, Lipson Hod. How transferable are features in deep neural networks? Adv Neural Inf Process Syst 2014;27. URL https://dl.acm.org/doi/10.5555/2969033.2969197. Retrieved on April 19, 2024.

[8] Snell Jake, Swersky Kevin, Zemel Richard. Prototypical networks for few-shot learning. In: Proceedings of the 31st international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 4080–90, URL https://dl.acm.org/doi/10.5555/3294996.3295163. Retrieved on April 19, 2024.

[9] Sun Qianru, Liu Yaoyao, Chua Tat-Seng, Schiele Bernt. Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 403–12. http://dx.doi.org/10.1109/CVPR.2019.00049, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00049. Retrieved on April 19, 2024.

[10] Tian Yonglong, Wang Yue, Krishnan Dilip, Tenenbaum Joshua B, Isola Phillip. Rethinking few-shot image classification: A good embedding is all you need? In: Computer vision – ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XIV. Berlin, Heidelberg: Springer-Verlag; 2020, p. 266–82. http://dx.doi.org/10.1007/978-3-030-58568-6_16, Retrieved on April 19, 2024.

[11] Mahajan Kushagra, Sharma Monika, Vig Lovekesh. Meta-DermDiagnosis: Few-shot skin disease identification using meta-learning. In: 2020 IEEE/CVF conference on computer vision and pattern recognition workshops. IEEE; 2020, http://dx.doi.org/10.1109/cvprw50498.2020.00373, Retrieved on April 19, 2024.

[12] Deuschel Jessica, Firmbach Daniel, Geppert Carol I, Eckstein Markus, Hartmann Arndt, Bruns Volker, et al. Multi-prototype few-shot learning in histopathology. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 620–8. http://dx.doi.org/10.1109/ICCVW54120.2021.00075, Retrieved on April 19, 2024.

[13] Parvatikar Akash, Choudhary Om, Ramanathan Arvind, Jenkins Rebekah, Navolotskaia Olga, Carter Gloria, et al. Prototypical models for classifying high-risk atypical breast lesions. In: Medical image computing and computer assisted intervention–mICCAI 2021: 24th international conference, strasbourg, France, September 27–October 1, 2021, proceedings, part VIII 24. Springer; 2021, p. 143–52. http://dx.doi.org/10.1007/978-3-030-87237-3_14, Retrieved on April 19, 2024.

[14] Yan Jin, Feng Kaiyuan, Zhao Hongyu, Sheng Kai. Siamese-prototypical network with data augmentation pre-training for few-shot medical image classification. In: 2022 2nd international conference on frontiers of electronics, information and computation technologies. IEEE; 2022, p. 387–91. http://dx.doi.org/10.1109/ICFEICT57213.2022.00075, Retrieved on April 19, 2024.

[15] Prabhu Viraj, Kannan Anitha, Ravuri Murali, Chablani Manish, Sontag David, Amatriain Xavier. Few-shot learning for dermatological disease diagnosis. In: Meta learning with medical imaging and health informatics applications. Elsevier; 2023, p. 235–52. http://dx.doi.org/10.1016/b978-0-32-399851-2.00022-3, Retrieved on April 19, 2024.

[16] Thrun Sebastian, Pratt Lorien. Learning to learn: Introduction and overview. In: Learning to learn. Springer US; 1998, p. 3–17. http://dx.doi.org/10.1007/978-1-4615-5529-2_1, Retrieved on April 19, 2024.

[17] Vanschoren Joaquin. Meta-learning: A survey. 2018, http://dx.doi.org/10.48550/arXiv.1810.03548, arXiv e-prints, arXiv:1810.03548. Retrieved on April 19, 2024.

[18] Hospedales Timothy, Antoniou Antreas, Micaelli Paul, Storkey Amos. Meta-learning in neural networks: A survey. IEEE Trans Pattern Anal Mach Intell 2022;44(9):5149–69. http://dx.doi.org/10.1109/TPAMI.2021.3079209, Retrieved on April 19, 2024.

[19] Parnami Archit, Lee Minwoo. Learning from few examples: A summary of approaches to few-shot learning. 2022, http://dx.doi.org/10.48550/arXiv.2203.04291, arXiv e-prints, arXiv:2203.04291. Retrieved on April 19, 2024.

[20] Finn Chelsea, Abbeel Pieter, Levine Sergey. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th international conference on machine learning, vol. 70. JMLR.org; 2017, p. 1126–35, URL https://dl.acm.org/doi/10.5555/3305381.3305498. Retrieved on April 19, 2024.

[21] Mishra Nikhil, Rohaninejad Mostafa, Chen Xi, Abbeel Pieter. A simple neural attentive meta-learner. In: International conference on learning representations. 2018, URL https://openreview.net/forum?id=B1DmUzWAW. Retrieved on April 19, 2024.

[22] Koch Gregory, Zemel Richard, Salakhutdinov Ruslan, et al. Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop, vol. 2. (1). Lille; 2015, URL https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf. Retrieved on April 19, 2024.

[23] Vinyals Oriol, Blundell Charles, Lillicrap Timothy, Kavukcuoglu Koray, Wierstra Daan. Matching networks for one shot learning. In: Proceedings of the 30th international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2016, p. 3637–45, URL https://dl.acm.org/doi/10.5555/3157382.3157504. Retrieved on April 19, 2024.

[24] Sung Flood, Yang Yongxin, Zhang Li, Xiang Tao, Torr Philip HS, Hospedales Timothy M. Learning to compare: Relation network for few-shot learning. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. 2018, p. 1199–208. http://dx.doi.org/10.1109/CVPR.2018.00131, URL https://doi.org/10.1109/CVPR.2018.00131. Retrieved on April 19, 2024.

[25] Motiian Saeid, Jones Quinn, Iranmanesh Seyed Mehdi, Doretto Gianfranco. Few-shot adversarial domain adaptation. In: Proceedings of the 31st international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 6673–83, URL https://dl.acm.org/doi/10.5555/3295222.3295412. Retrieved on April 19, 2024.

[26] Sahoo Doyen, Le Hung, Liu Chenghao, Hoi Steven CH. Meta-learning with domain adaptation for few-shot learning under domain shift. 2019, URL https://openreview.net/forum?id=ByGOuo0cYm. Retrieved on April 19, 2024.

[27] Chen Xiao. Enhancing prototypical networks for few-shot learning. 2024, URL https://web.stanford.edu/~markcx/sample-project/CS231N_project_metaLearning.pdf. Retrieved on April 19, 2024.

[28] Laenen Steinar, Bertinetto Luca. On episodes, prototypical networks, and few-shot learning. In: Beygelzimer A, Dauphin Y, Liang P, Vaughan J Wortman, editors. Advances in neural information processing systems. 2021, URL https://openreview.net/forum?id=bJaZ8leI0QJ. Retrieved on April 19, 2024.

[29] Liu Xue-Jun, Li Kai-li, Luan Hai-ying, Wang Wen-hui, Chen Zhao-yu. Few-shot learning for skin lesion image classification. Multimedia Tools Appl 2022;81(4):4979–90. http://dx.doi.org/10.1007/s11042-021-11472-0, Retrieved on April 19, 2024.

[30] Singh Rishav, Bharti Vandana, Purohit Vishal, Kumar Abhinav, Singh Amit Kumar, Singh Sanjay Kumar. MetaMed: Few-shot medical image classification using gradient-based meta-learning. Pattern Recognit 2021;120:108111. http://dx.doi.org/10.1016/j.patcog.2021.108111, Retrieved on April 19, 2024.

[31] Tschandl Philipp, Rosendahl Cliff, Kittler Harald. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5(1). http://dx.doi.org/10.1038/sdata.2018.161, Retrieved on April 19, 2024.

[32] Combalia Marc, Codella Noel CF, Rotemberg Veronica, Helba Brian, Vilaplana Veronica, Reiter Ofer, et al. BCN20000: Dermoscopic lesions in the wild. 2019, http://dx.doi.org/10.48550/arXiv.1908.02288, arXiv e-prints, arXiv:1908.02288. Retrieved on April 19, 2024.

[33] Cassidy Bill, Kendrick Connah, Brodzicki Andrzej, Jaworek-Korjakowska Joanna, Yap Moi Hoon. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Med Image Anal 2022;75:102305. http://dx.doi.org/10.1016/j.media.2021.102305, Retrieved on April 19, 2024.

[34] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition. 2016, p. 770–8. http://dx.doi.org/10.1109/CVPR.2016.90, Retrieved on April 19, 2024.

[35] Chamarthi Sireesha, Fogelberg Katharina, Brinker Titus J, Niebling Julia. Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods. Inform Med Unlocked 2024;44:101430. http://dx.doi.org/10.1016/j.imu.2023.101430, Retrieved on April 19, 2024.

[36] Ji Zhong, Chai Xingliang, Yu Yunlong, Pang Yanwei, Zhang Zhongfei. Improved prototypical networks for few-shot learning. Pattern Recognit Lett 2020;140:81–7. http://dx.doi.org/10.1016/j.patrec.2020.07.015, Retrieved on April 19, 2024.