Medical Imaging Benchmark for Few-Shot Image Classification

Lourenço Tourais, Instituto Superior Técnico

***Abstract*—Nowadays machines are not able to learn and generalize quickly from few examples like humans can. This difference between humans and machines and the existence of several tasks where there are hardly any available examples lead to the increasing research popularity on how to learn from few examples. This work focuses on the few-shot medical imaging classification area that has the intent to classify new medical images with just a few training images. This thesis proposes the fsMIB, a few-shot learning benchmark composed of several public medical imaging datasets that can drive the future evaluation of the state-of-the-art progress in few-shot image classification models towards the goal of discovering models that allow a better classification with few samples in the medical context, making possible the future design of reliable medical imaging diagnostic systems. The results obtained by evaluating few-shot metric learning models on the proposed benchmark are, in general, consistent with those obtained in the Meta-Dataset, the currently used benchmark. Training the models in the Meta-Dataset datasets achieved better results than training the models in some of the medical datasets of the fsMIB, when evaluating the results on the other available medical datasets. Finally, it is concluded that the current approach to evaluate FSL models may not be suitable to accurately assess the effectiveness of these models in medical imaging classification.**

***Index Terms*—Few-shot Learning, Medical Image Classification, Few-shot Benchmark, Deep Learning.**

I. Introduction

THE early recognition of diseases, for instance, cancer or cardiovascular diseases, can save lives and greatly improve patient outcomes and quality of life. Medical imaging techniques, like x-ray radiography, magnetic resonance imaging and ultrasound, are frequently used for diagnosis purposes. Nonetheless, the analysis of these images is often subjective, as different clinicians may interpret the same images differently, based on their different levels of experience. Furthermore, the learning curve to analyse medical images may lead an observer to apply a different score to the same image than the applied in a previous reading [1]. This subjectivity can lead to poor reproducibility and low accuracy, especially for less experienced clinicians. For example, the accuracy of experienced dermatologists in examining skin lesions is 75% - 84%, while the accuracy of inexperienced clinicians is even lower [2]. Besides, there is a shortage of experienced clinicians in many parts of the world.

Recent advancements in Deep Neural Networks (DNNs) have shown promise in creating automatic diagnosis systems that may surpass the diagnostic accuracy of human clinicians [3]. These systems may provide objective answers and be available in areas where there are not enough clinicians. Notwithstanding, "there is a large consent that successful training of deep networks requires many thousand annotated training samples" [4] and the use of Graphical Processing Units (GPUs), which can be difficult to obtain and come with significant energy expenditure and carbon emissions [5]. To address these issues, Few-Shot Learning (FSL), a subfield of Machine Learning (ML), aims to enable machines to reproduce and enhance the human ability to learn new concepts from a small amount of data.

In recent years, there have been many advances in FSL, with a variety of models [6]–[18] proposed in the literature. However, most of these models have mainly been evaluated on benchmarks composed of natural and hand-made images [6], [19], [20], which can limit their applicability to other domains. Among these benchmarks, Meta-Dataset [20] is currently the most widely used for evaluating new few-shot image classification models. Notably, there has been limited research on applying FSL in the context of medical imaging classification, where "an abundance of well-labeled data (...) is desirable but rarely available" [21] due to the high costs of human annotation, the need for well-trained clinicians to annotate medical images and the prevalence of rare diseases.

According to a review of the literature, it is clear that "in medical imaging, few-shot learning research is sparse, limited to private data sets and is at its early stage." [22]. In order to advance this field, it is necessary to propose multiple benchmarks to assess progress in different directions. Consequently, this work proposes a new benchmark called Few-shot Medical Imaging Classification Benchmark (fsMIB) that aims to enable the study of FSL in the field of medical imaging analysis. The ultimate goal of this benchmark is to enable the assessment of future few-shot models on medical images, rather than just natural and hand-made images, as it is currently possible with the Meta-Dataset benchmark [20]. In line with the ideas presented in the Meta-Dataset, the fsMIB intends to be a publicly available benchmark that is easy to use and covers a variety of medical domains. Finally, the performance of five few-shot metric learning models is compared using the fsMIB. This comparison permit the verification of whether the performance of the models on the fsMIB is consistent with their performance on the Meta-Dataset. It also allows for the verification of the effectiveness of the new benchmark as a means of evaluating the performance of future few-shot image classification models.

II. Related Work

The FSL problem has been tackled through several approaches, which can be divided in two main groups:

optimization-based methods and metric-based methods. Optimization-based methods address the problem by optimizing the initialization of a model, as a way to have a better starting point for fine-tuning on the support set with a small number of steps, e.g., MAML [9] or by optimizing the optimizer in order to learn an optimized update rule that allows the model to converge with a small number of gradient steps and a small amount of training data in the support set, e.g., Meta-Learner LSTM [8]. Metric-based methods learn a representation space, in which images from an episode can be represented, and learn how to compare the embeddings from the query set to the ones from the support set, so that they can assign to a query image the class of the support set that has the most similar images, e.g., Matching Networks [6], Prototypical Networks (ProtoNet) [7] and Relation Networks [23]. The similarity measurement is based on a distance metric. The referred models are the earliest models adopted to address the FSL setting where the training and test examples came from a single domain, e.g., omniglot [19] and mini-imagenet [6]. Nevertheless, these models have performed poorly in the challenging scenario of cross-domain few-shot image classification.

Contrarily, when evaluated on the Meta-Dataset, the current few-shot image classification benchmark designed to enable the study of cross-domain FSL, recent models have shown promising results. The state-of-the-art in few-shot image classification, as determined by this benchmark[1], is the Cross-Domain Few-Shot Learning with Task-Specific Adapters (TSA) model [18]. This model is based on the Universal Representation Learning from Multiple Domains for Few-Shot Classification (URL) model [17], which also performs well on the benchmark. Both models learn one domain-specific feature extractors for each training dataset and benefit from the knowledge distillation [24], [25] process proposed by URL, which consolidates the knowledge from the multi-domain feature extractors into a single feature extractor. Other models, such as SUR [10] and URT [11], include attention mechanisms for selecting features from the domain-specific feature extractors, but are computationally more expensive in comparison to URL and TSA.

Several recent few-shot learning models have been proposed to address the challenge of adapting a feature extractor to unseen domains. These models, including Conditional Neural Adaptive Processes (CNAPS) [12], Simple CNAPS (SCNAPS) [15], Transductive CNAPS (TCNAPS) [16], Few-shot Learning with a Universal Template (FLUTE) [14], Multi-Mode Modulator (Tri-M) [13], and TSA [18] introduce special layers with a small number of parameters that adapt the feature extractor to the data of the support set. Except TSA, all these models use Feature-wise Linear Modulation (FiLM) [26] layers to condition the feature extractor to the task at hand. Then, they employ an auxiliary network that, conditioned on a representation of the support set, outputs the FiLM parameters Differently, TSA learns residual adapters [27] parameters directly from the support set. CNAPS, SCNAPS, and TCNAPS differ in the type of classifier they use: CNAPS uses an adaptive parametric classifier, SCNAPS uses a non-parametric classifier and TCNAPS improves SCNAPS classifier by incorporating the knowledge of the unlabeled instances from the query set.

Recently, there has been an increase in research on the application of FSL to problems containing medical images. Guo et al. [28] introduced the Broader Study of Cross-Domain Few-Shot Learning (BSCD-FSL) benchmark, which examines the application of FSL methods to a spectrum of images with different degrees of similarity compared to natural images, i.e., satellite images, dermatology images and radiological images. Similarly, Shakeri et al. [22] proposed the Few-shot Classification of Histological Images (FHIST) benchmark, which focuses specifically with histology images. What is more, the BSCD-FSL benchmark is not specifically designed for medical imaging, even though it contains medical imaging datasets. Both benchmarks include four datasets and its findings are consistent with each other. Further than the proposal of benchmarks, there have been previous studies [29], [30] that have applied FSL models to specific medical applications, such as Covid-19 diagnosis [30].

III. Background

A. Meta-Dataset: Few-Shot Classification Benchmark

Meta-Dataset [20] is the current few-shot image classification benchmark proposed as a new benchmark that aims to provide a more realistic, heterogeneous, large-scale, and diverse environment for training, testing, and evaluating few-shot learners, as well as to facilitate the analysis of cross-domain generalization. It also allows to analyse the impact on model performance of more data, heterogeneous training sources, pre-trained weights, and meta-training.

Meta-Dataset is composed of 10 datasets: ILSVRC-2012 (ImageNet) [31], Omniglot [19], Aircraft [32], CUB-200-2011 (Birds) [33], Describable Textures (DTD) [34], Quick Draw [35], Fungi [36], VGG Flower [37], Traffic Signs [38] and MSCOCO [39]. Each of these datasets was chosen given that: it was free and easy to obtain, it spanned a variety of visual concepts (natural and human-made) and it varied in how fine-grained the class definition is.

Meta-Dataset improves the previous benchmarks, thanks to having a significantly larger amount of data and being composed by multiple datasets. Also, it allows research on how different sources of data may be explored to improve few-shot learners generalization. Meta-Dataset has other constraints: each episode is drawn from a single dataset to ensure the similarity to real classification problems; the Traffic Signs and MSCOCO datasets are reserved for evaluating model performance; and the remaining datasets contribute, approximately, 70% / 15% / 15% of their classes to the training, validation and test splits, respectively.

Meta-Dataset's Episode Sampling Algorithm Meta-Dataset [20] proposes an algorithm to sample episodes maintaining a realistic class imbalance in terms of shots and ways. The proposed steps for sampling episodes for a given split are:

[1]The performance of the few-shot models applied to the Meta-Dataset benchmark is presented in: https://github.com/google-research/meta-dataset/. (Accessed in 11-01-2023)

0) Select a dataset $\mathcal{D}$ at random, 1) Sample a set of classes $\mathcal{C}$ from the classes of $\mathcal{D}$ assigned to the requested split, and 2) sample support and query examples from $\mathcal{C}$.

Average Rank Metric Meta-Dataset authors developed a metric called average rank that encompasses in a single value the performances of a model on all the datasets it was evaluated.

For each dataset, the models are ranked in decreasing order of accuracies. If the differences between the models are not statistically significant, the models are considered to be tied. The tied models are assigned with the same rank, which is the average of the ranks they would have without the ties. Finally, each model has a rank for each dataset and this rank is averaged across the datasets to determine the performance of each model in the full benchmark

B. FiLM

A FiLM layer [26] consists of an affine map that scales and shifts the output channels $\mathbf{f}$ of a NN layer using two parameters, γ and β, as shown in Equation 1:

$$FiLM(\mathbf{f}, \gamma, \beta) = \gamma \odot \mathbf{f} + \beta \tag{1}$$

where $\gamma \in \mathbb{R}^C$ and $\beta \in \mathbb{R}^C$ are vectors containing one scalar for each output channel.

C. Residual Adapters

A Residual Adapter Layer [27] consists of mapping the channels $\mathbf{f}$ of a residual connection through matrix multiplication. Mathematically, it is defined as:

$$r_\alpha(\mathbf{f}) = \mathbf{f} \times \alpha, \tag{2}$$

with $\alpha \in \mathbb{R}^{C \times C}$. These adapters are similar to the FiLM conditioning method [26], but the number of parameters to be learned is higher.

IV. Few-shot Classification: Task Formulation and Approaches

A few-shot image classification task has the objective to learn how to classify images based on a limited amount of labeled images. This task, also known as an episode, consists in two sets of images: the support set for training and the query set for testing. Consider an episode τ, the support set $\mathcal{S}^\tau = \{(x_k^\tau, y_k^\tau)\}_{k=1}^{|\mathcal{S}^\tau|}$ contains N classes and a small number of labeled examples per class (k_c per class with $c \in 1, ..., N$), with $|\mathcal{S}^\tau| = \sum_c k_c$, and the query set $\mathcal{Q}^\tau = \{(x_t^{\tau *}, y_t^{\tau *})\}_{t=1}^{|\mathcal{Q}^\tau|}$ contains only $|\mathcal{Q}^\tau|$ unlabeled examples to be classified. Each example (x, y) is composed of an image and its corresponding label. The model must learn through the support set data, to predict the labels of the query set and to evaluate the accuracy of the predictions. In the literature, the term *shot* refers to the number of support examples per class and *way* refers to the number of classes present in each task. When the support set is balanced ($k_c = k,\ \forall c$), the classification task is described as 'N-way, k-shot'. Each episode for evaluation is sampled from a larger set $\mathcal{D}_{test}$ by sampling N classes with the desired number of examples per class and by sampling $|\mathcal{Q}^\tau|$ unlabeled examples.

Before addressing a Few-shot image classification task, it is necessary to embed prior knowledge into the model by training it in a, typically large set $\mathcal{D}_{train}$, turning the model into a general learner capable of fast adaptation to different contexts. Subsequently, $\mathcal{D}_{test}$ is used to evaluate the model's ability to learn from few samples. There are two different concepts of training in FSL: training through $\mathcal{D}_{train}$ to learn a general model and training through the support set of an episode sampled from $\mathcal{D}_{test}$ to classify the correspondent query set. With the aim of avoiding confusion, the term *training* will refer to the process of learning from $\mathcal{D}_{train}$ and the term *testing* will refer to the process of learning from episodes sampled from $\mathcal{D}_{test}$.

There is no specific procedure on how to exploit $\mathcal{D}_{train}$. In the literature, there are two approaches: training episodically or training in a standard supervised learning manner. Initially, it was believed that training episodically was the most promising choice to exploit $\mathcal{D}_{train}$. Nevertheless, recent research [40]–[43] have reported that the performance of non-episodic approaches can be comparable or even better than episodic approaches, especially when there is a large gap between source and target domains. Training episodically, as proposed by Vinyals et al. [6] involves matching the training and testing conditions. Thus, as a way to enable fast learning, training is performed by sampling different episodes from $\mathcal{D}_{train}$, as done in the testing phase. The model learns from the support sets of these episodes and classifies their corresponding query sets. This process is intimately related to the concept of "learning-to-learn", known as meta-learning. During the training phase, the model learns to learn from the support set with the intent to classify the query set. This is repeated until the model is able to learn from any task defined by a support set.

The non episodic approach consists in training a feature extractor, for instance, with a linear classifier, in a standard supervised learning. This leads the model to represent images with an embedding in a meaningful representation space, which can potentially be used for images of classes that were not present in $\mathcal{D}_{train}$. The algorithm for performing few-shot classification in the non-episodic approach varies depending on the specific model being used.

A. Few-Shot Metric Learning Models

ProtoNet Prototypical Networks (ProtoNet) [7] compute an embedding, known as prototype, for each class. A prototype is the mean of the embeddings of each class support samples. Each query point is classified with the class whose prototype is nearest to it. The probability of a query image $\mathbf{x}$ belonging to a particular class k is estimated as shown in Equation 3:

$$p(y^* = k|\mathbf{x}^*, \mathcal{S}^\tau) = \text{softmax}(-d(f_\phi(\mathbf{x}^*), \mathbf{c}_k)) \tag{3}$$

where d is the Euclidean distance and ϕ represents the parameters of the feature extractor f.

SCNAPS Simple CNAPS (SCNAPS) [15] constructs predictive distributions, given an input $\mathbf{x}^*$, as:

$$p(y^*|\mathbf{x}^*, \theta, \mathcal{S}^\tau) = p(y^*|\mathbf{x}^*, \theta, \psi^\tau = \psi_\phi(\mathcal{S}^\tau)) \tag{4}$$

where θ are global classifier parameters and ψ^τ are episode-specific parameters, produced by a function ψ_ϕ that acts on the support set $\mathcal{S}^\tau$. ψ_ϕ is parameterized by another set of parameters called adaptation network parameters, denoted by ϕ, which together with θ are the model's learnable parameters.

The global classifier parameters parameterize a feature extractor, denoted as $f_\theta(x)$. The episode-specific parameters, represented by ψ^τ, are then used to adapt the feature representations obtained with $f_\theta(\cdot)$ through the use of FiLM layers. $f_\theta(\cdot)$ denotes the *unadapted* feature extractor and $f_\theta(\cdot; \psi^\tau)$ the adapted feature extractor for task τ, where ψ^τ are the set of FiLM parameters inserted into the feature extractor.

An auxiliary network called adaptation network is trained on multiple tasks to produce the FiLM parameters, given a support set $\mathcal{S}^\tau$. Once the FiLM parameters are produced, the feature extractor is adapted to the current episode and is denoted as f^τ_θ.

SCNAPS uses a deterministic classifier that computes the Mahalanobis distance relative to each class k by estimating a mean $\boldsymbol{\mu}_k$ and regularized covariance $\mathbf{Q}_k$ in the adapted feature space. Finally, SCNAPS computes the class probabilities as:

$$p(y^* = k|\mathbf{z}^*, \mathcal{S}^\tau) = \text{softmax}(-d_k(\mathbf{z}^*, \boldsymbol{\mu}_k)) \tag{5}$$

using a deterministic, fixed Mahalanobis distance, d_k:

$$d_k(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})^T (\mathbf{Q}^\tau_k)^{-1}(\mathbf{x} - \mathbf{y}). \tag{6}$$

TCNAPS Transductive CNAPS (TCNAPS) [16] is an extension of SCNAPS [15] to the transductive FSL setting. It improves test-time classification accuracy using unlabelled data by adopting a regularized Mahalanobis-distance-based soft k-means clustering procedure.

In TCNAPS, the adaptation network of SCNAPS is extended to incorporate both support set and query set information to produce FiLM parameters adapted to the task at hand.

TCNAPS estimates the classes of the unlabelled data in the query set and uses them to refine parameters $\boldsymbol{\mu}_k$ and $\mathbf{Q}_k$. A soft k-means procedure is introduced for the updates. This procedure initializes the weights using the labeled support set, uses the weights to compute the query examples classes and then updates the weights using both the support and query sets. To conclude, it is repeated iteratively until reaching a maximum number of iterations or until the query set class assignments stop changing.

URL The Universal Representation Learning from Multiple Domains for Few-shot Classification (URL) proposes a two stage procedure for learning multi-domain representations. In the first stage individual domain-specific networks are trained for each of the training datasets. Each of these networks is composed of a specific feature extractor and classifier with their own set of parameters. In the second stage, a single multi-domain network is obtained through the process of knowledge distillation [24], [25], which transfers the knowledge of the trained individual domain-specific networks to a single model.

Knowledge Distillation is performed at the prediction and feature levels by minimizing the distance between the predictions of the multi-domain and corresponding single-domain network and between the multi-domain and single-domain features for given training examples.

The final model is obtained by combining a universal feature extractor f_ϕ with a task-specific classifier c_ϑ, where ϑ are the parameters of the classifier. The feature extractor is used to embed the support and query images for the purpose of classification. These embeddings are then transformed through a linear transformation A_ϑ, with learnable parameters ϑ, to align the extracted features with the target task. The ProtoNet classification mechanism is then used to compute prototypes and estimate the likelihood of a query sample $\mathbf{x}^*$ belonging to a particular class, using the negative cosine similarity as the distance metric for Equation 3. This classifier is known as Nearest Centroid Classifier (NCC).

TSA The Cross-domain Few-shot Learning with Task-specific Adapters (TSA) [18] is a cross-domain FSL model that is capable of adapting to unseen tasks even when there is a large domain gap between the training and test sets. This model was developed by the same authors of the URL [17] model and benefits from the same feature extractor. Unlike previous approaches [12]–[14], TSA learns task-specific weights directly from the support set, rather than relying on an auxiliary network to dynamically learn these weights.

TSA adds task-specific weights to each residual network block of a ResNet [44] backbone, through the use of residual adapters. The residual adapters weights are estimated directly from a small support set. These adapters intend to improve the representations of the task-agnostic feature extractor.

The classifier adaptation strategy employed by TSA is the same as that used in URL. Therefore, the final model is represented by a NCC classifier that is applied to the adapted feature extractor combined with a pre-classifier transformation.

V. fsMIB: Few-Shot Medical Imaging Classification Benchmark

The fsMIB is a new benchmark that has the objective to offer an environment for measuring progress in few-shot classification in the field of medical imaging analysis. It is composed of multiple datasets, with a focus on radiological datasets as they are composed of 2D images and there are public radiological datasets with a large number of correctly labeled x-ray images. In addition to the radiological datasets, fsMIB also includes two additional datasets from different medical imaging domains: ultrasound and dermatoscopy. This enables the study of cross-domain FSL.

In line with the characteristics of Meta-Dataset [20], the datasets in fsMIB are public and freely available for non-commercial research purposes and easily convertible to the

format of the Meta-Dataset reader. They are composed of easily manageable images that include a variety of medical imaging types. In more detail, the fsMIB include: different dataset sizes in terms of the number of images, since FSL models are not necessarily trained on small sample sizes and having a range of dataset sizes allows for the analysis of training with different numbers of images; diversified medical imaging techniques to allow the study of cross-domain scenario; and radiological datasets with different applications, e.g., covid versus musculoskeletal abnormality, to measure the generalization ability of models in the in-domain scenario.

Unlike other FSL benchmarks, where the objective is to train a model on a set of classes and evaluate on the remaining classes, fsMIB is designed for medical imaging and takes into account the specific characteristics of medical classification tasks, which have a fixed number of classes. Specifically, the number of ways (i.e., the number of classes in an episode) is fixed and equal to the number of classes in the medical classification task. This means that each dataset can only be used for training or evaluation, rather than both, to ensure that the classification tasks are different in training and evaluation. This design allows the evaluation of FSL models on real-world medical imaging tasks.

A. *fsMIB's datasets*

The fsMIB contains multiple datasets, each with a different classification task. Each dataset has more information in the Section 4.4 from the Master's thesis [45], where this work is based.

The datasets in fsMIB are:

- **covid dataset**: This dataset leverages data from the Covid-QU-Ex[2] [46] dataset. Its task is to classify chest x-ray images as being from a patient with covid, a patient with other pathologies or a normal patient.
- **breast dataset**: This dataset is based on the Breast Ultrasound Images[3] [47] dataset and has the purpose to classify breast ultrasound images as benign, malignant or normal, according to the presence or absence of cancer.
- **elbow, finger, forearm, hand, humerus, shoulder and wrist datasets**: These datasets were extracted from the MURA[4] [48] dataset. Each of these datasets classification task involves labeling its images as having or not having a musculoskeletal abnormality.
- **tuberculosis dataset**: This dataset contains chest x-ray images and holds the task to classify each image as manifesting or not manifesting tuberculosis. This dataset leverages data from the Shenzhen set[5] [49].
- **skinlesion dataset**: This dataset's task is to classify the dermatoscopy images of skin lesions into one of the following classes: Actinic Keratoses and Intraepithelial Carcinoma (akiec); Basal Cell Carcinoma (bcc); Benign Keratosis (bkl); Dermatofibroma (df); Melanoma (mel); Melanocytic Nevi (nv) and Vascular Skin Lesions (vasc). This dataset data is taken from the ISIC 2018[6] [50], [51] dataset.

[2]https://www.kaggle.com/datasets/anasmohammedtahir/covidqu (Accessed in 13-01-2023)

[3]https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset (Accessed in 13-01-2023)

[4]https://stanfordmlgroup.github.io/competitions/mura/ (Accessed in 13-01-2023)

[5]https://openi.nlm.nih.gov/faq#faq-tb-coll (Accessed in 13-01-2023) – The dataset is mentioned in "More Questions → Data Collection → I have heard about the Tuberculosis collection. Where can I get those images?" of the linked website.

VI. Experiments

The performance of the few-shot metric learning models described in Section IV-A was evaluated with the following experiments:

- **meta-med** - apply the models trained on the Meta-Dataset to random episodes sampled from **all** the fsMIB datasets and report the results. This experiment tests the generalization of the models from natural and hand-made images (as used in the Meta-Dataset) to medical images;
- **med-med** - train the models in **elbow, forearm, hand, humerus, tuberculosis** and **wrist** datasets of the fsMIB and evaluate them in **breast, covid, finger, shoulder** and **skinlesion**. This experiment tests the generalization of the models from one set of medical images to another set of medical images.

For the **med-med** experiment, the datasets included only x-ray datasets in the training split. This was done to test generalization: from a radiological task to another different radiological task and from a radiological task to a medical imaging task, such as ultrasound or dermatoscopy. As a result, **breast** and **skinlesion** were necessarily assigned to the testing split. The other datasets were split into the training and testing partitions, with **covid** and **tuberculosis**, each being placed in one partition, and **finger** and **shoulder**, from the MURA datasets, both being placed in the testing partitions. This was done in order to include a range of granularities in the testing split, with the smaller bones in the **finger** dataset being contrasted with the larger ones in the **shoulder** dataset, while leaving the other MURA datasets for the training step. Albeit this experiment is composed as described, the fsMIB creates new opportunities for studying FSL by conducting experiments with different datasets in the training and testing partitions.

VII. Training and Testing Details

The few-shot metric learning models studied used an adapted ResNet-18 with an output dimension of 512 channels as a backbone for the feature extractor. This ResNet-18 was pretrained on the training split of the ImageNet subset of the Meta-Dataset for the ProtoNet, SCNAPS and TSCNAPS models. The URL and TSA models did not undergo this pretraining stage, as they train a separate network for each of the training datasets. The meta-med experiment used the checkpoints of the models trained on the Meta-Dataset, as provided by the authors of the few-shot classification models. In addition, the experiments were evaluated for tasks with 1, 5, 15, 30 and 50 shots per class in the support set and a fixed number of 15 query images per class. The training of the

[6]https://challenge.isic-archive.com/landing/2018/47/ (Accessed in 13-01-2023)

models on the fsMIB used the Meta-Dataset episode sampling algorithm to generate episodes with variable shots per class, creating the opportunity to train the models only once, while still subjecting them to different shots.

Since the objective of the experiments is to assess the performance of these models, it was decided to use the literature established hyperparameters for training, rather than tuning the hyperparameters to improve the results. Consequently, to maintain coherence with the literature, the images in this work were downsized to a resolution of 84×84 through shrinking without padding. This reduction in size can compromise the quality of medical images with higher resolutions, as it may distort the images and cause relevant features to be lost. Alternatively, using a higher resolution, e.g., 224×224, may have been a better option. Despite that, the available checkpoints trained on the Meta-Dataset were based on a ResNet-18 adapted for small resolutions and it was not possible to meet the memory and time requirements for higher resolutions.

To conclude, a random classifier was implemented as a random version of ProtoNet using randomly initialized parameters for the feature extractor and resulting in the generation of random embeddings of the images. This random baseline was not trained and was only used for testing. The classification of the query images used the non-parametric classifier of ProtoNet. This approach was chosen as a baseline, given that it is challenging for a machine to accurately classify images with only a small number of labeled examples.

A. Meta-Med Results

The results of this experiment are shown in Figures 1 and 2. Both figures contain the same information, however Figure 1 allows a direct comparison of model performance per dataset, while Figure 2 allows for a direct comparison of dataset performance per model.

Upon examining the figures, it is evident that the **breast, covid, skinlesion** and **tuberculosis** datasets are learned by the models, as there is a significant difference in accuracy compared to the random classifier. On the contrary, when applied to the MURA datasets, the models struggle to outperform random guessing. Despite this, **forearm, humerus** and **shoulder** show slight improvements when the number of shots available is increased. The complexity of the MURA datasets can be attributed to the images being very different from each other, the semantic information present in each image being limited and the image resolution of 84×84 removing relevant information for the classification tasks. In conclusion, these results show that the MURA datasets are more difficult to learn compared to the non-MURA ones. This is clear in the plots in Figure 2.

Moreover, while the datasets **breast, covid** and **skinlesion** have significantly higher accuracies compared to random guessing, for 1 shot, the other datasets have not. In contrast, this is not the case for the other datasets. Nevertheless, this behaviour changes when more shots are available. These results align with what has been previously shown in the literature [6], [7], [20], since they demonstrate that the accuracy improves as the number of shots increases, with the greatest improvement occurring between 1 and 5 shots. In general, as more shots are provided, the rate of improvement decreases.

Besides the data shown in Figures 1 and 2, Table A.1, in the Appendix of the Master's thesis [45], contains the numerical results of the mean accuracies depicted in the plots, as well as the 95% confidence intervals.

B. Med-Med Results

The results of this experiment are presented in Figures 3 and 4, in the same way as in Section VII-A. The detailed numerical results are in Table A.2, in the Appendix of the Master's thesis [45]. Similar to **meta-med** findings, **breast, covid** and **skinlesion** are easier to learn and show a greater rate of accuracy improvement than **finger** and **shoulder** datasets. In spite of the difficulty of learning on the **shoulder** dataset, TCNAPS, the best model evaluated on this dataset, was able to achieve a mean accuracy 7% higher than the random classifier with just 15 shots. By contrast, for **finger**, the models struggled to perform better than random. As in the **meta-med** experiment, these results are in line with the literature [6], [7], [20], as they show that the accuracies increase, as the number of shots increases. Figure 3 shows that for the 1 shot scenario all the models start by randomly guessing the classes of query set images.

Despite the results being similar to those of **meta-med**, there are two unexpected results. The first is that **meta-med** results were better than **med-med** results, and this will be addressed in Section VII-C. The second is that URL performs poorly, achieving lower accuracies than ProtoNet and even less than random in the **covid** and **skinlesion** datasets. This issue will be addressed in Section VII-D. Even with poor results, URL bad performance did not prevent TSA, which is based on URL ideas, from achieving excellent results compared to the other models.

C. Comparing the Generalization Capability of Meta-Med and Med-Med Experiments

Meta-med and **med-med** experiments are proposed to evaluate the generalization capabilities of models trained on different datasets. The **meta-med** experiment used models trained on the Meta-Dataset to evaluate their performance on the fsMIB, whereas the **med-med** experiment trained and evaluated models on datasets from the fsMIB. As both the Meta-Dataset and fsMIB were proposed as benchmarks for cross-domain generalization, it is important to compare the results of these two experiments to determine which one demonstrates better generalization. The comparison must be limited to the **breast, covid, finger, shoulder** and **skinlesion** datasets, as these are the only datasets evaluated in the **med-med** experiment.

The results of the **meta-med** and **med-med** experiments are compared in Figure 5, which presents, for each dataset, a plot with the curves for each model showing the difference in mean accuracies between the two experiments. A positive difference indicates that **meta-med** generalizes better, while a negative difference indicates better performance of **med-med**. By looking into Figure 5, it is evident that **meta-med**

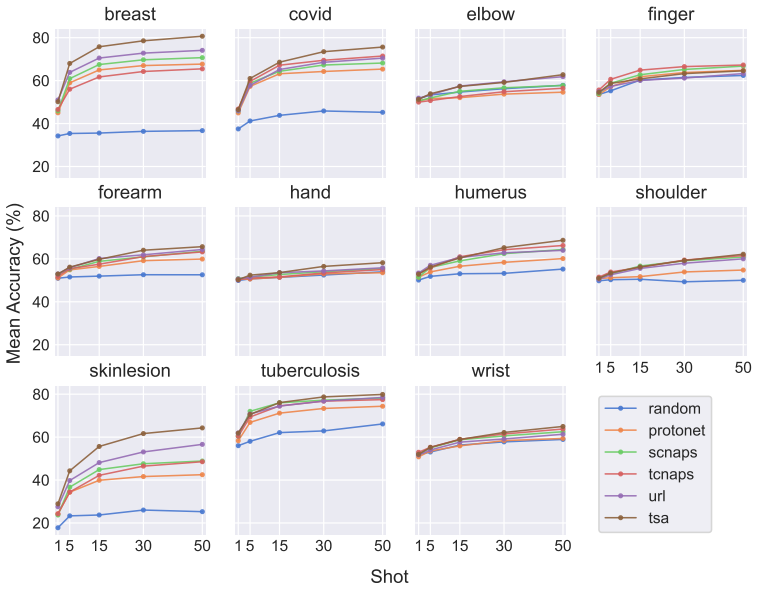
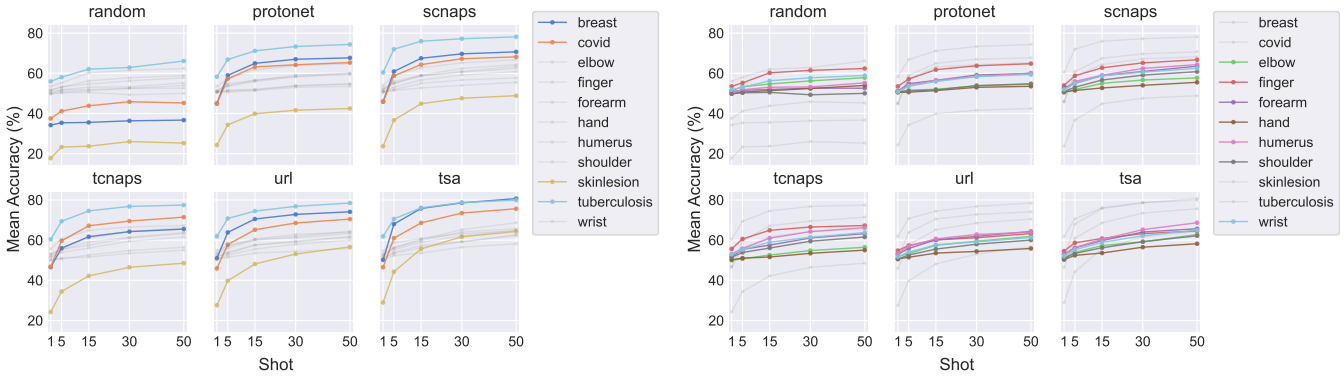


Fig. 1: Meta-med plots of the mean accuracies, in %, obtained by the few-shot models, given the number of shots, for each dataset.



(a) Plots highlighting non-MURA datasets.

(b) Plots highlighting MURA datasets.

Fig. 2: Meta-med plots of the mean accuracies, in %, obtained in the fsMIB datasets, given the number of shots, for each model. These plots were splitted in two to be more understandable, making more evident the differences between the classification of MURA and non-MURA datasets.

generalizes better. Specifically, for the datasets **breast, covid** and **skinlesion**, the difference in mean accuracies between the two experiments is higher than 10% for 5 to 50 shots. For the **shoulder** dataset, the differences, although smaller, tend to favor **meta-med**, since the differences are mostly on the positive side. Differently, the differences for the **finger** dataset are close to zero, indicating that this dataset is difficult to learn or may even be unlearnable under the conditions of the two experiments.

In the beginning, it was hypothesized that training models on medical images, more concretely in radiological images, would lead to improved generalization to other medical images, particularly other radiological images. Even so, the results presented in Figure 5 show that training on the Meta-Dataset leads to higher accuracies in the evaluation phase than training on the fsMIB. This suggests that the generalization is greater for Meta-Dataset trained models.

There are several potential explanations for this outcome. One possibility is that Meta-dataset contains a larger number of images than the fsMIB and its images are inherently more diverse, as each dataset represents a different type of image (e.g., birds or textures), with multiple classes within each type. Contrastingly, the fsMIB images are primarily x-rays, which are similar by nature, due to their black background and white bones. Additionally, the classes within each dataset of the fsMIB are restricted to medical diagnostic categories. The diversity within Meta-Dataset may enhance the feature extractor's learning of the underlying structure of an image, reducing the risk of overfitting the data.

Another possible explanation is the use of **elbow, forearm, hand, humerus, tuberculosis** and **wrist** as training datasets for the **med-med** experiment, given that five out of the six datasets are MURA datasets, which are known to be difficult to learn, as described in Sections VII-A and VII-B. Regardless of this, even if different datasets had been chosen, the issue pointed out in the first aforementioned explanation would still

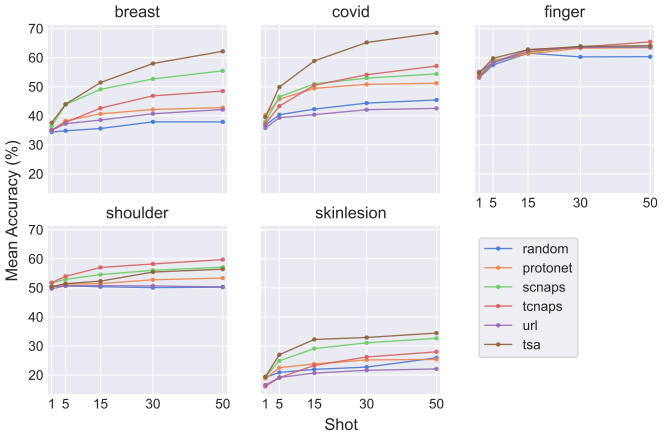


Fig. 3: **Med-med** plots of the mean accuracies, in %, obtained by the few-shot models, given the number of shots, for each dataset.

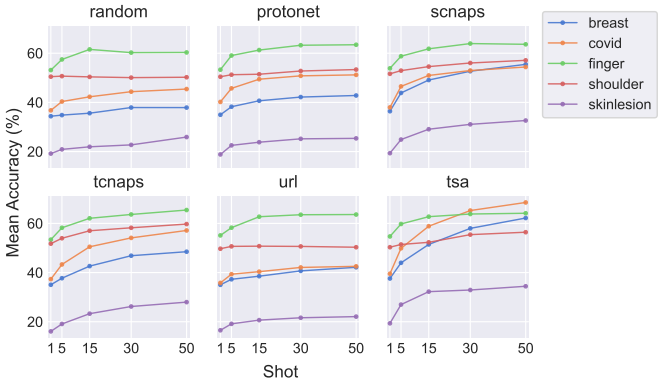


Fig. 4: **Med-med** plots of the mean accuracies, in %, obtained in the fsMIB datasets, given the number of shots, for each model.

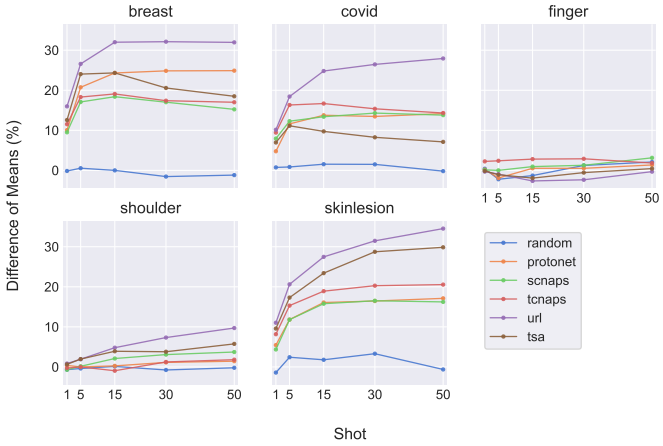


Fig. 5: Plots for each dataset of the difference of the mean accuracies, in %, between **meta-med** and **med-med** experiments, obtained by the few-shot models, given the number of shots.

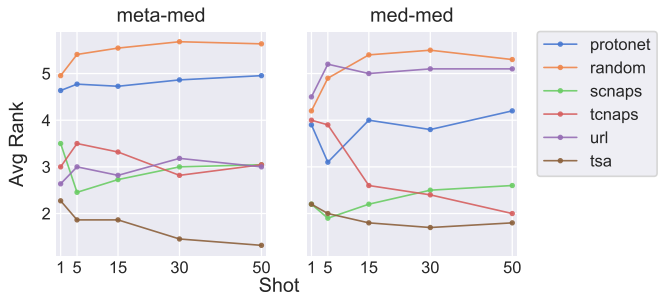


Fig. 6: Plots of the Average Rank, given the number of shots, for each of the few-shot image classification models and for each of the experiments.

exist.

D. Comparing the Capabilities of Few-Shot Metric Learning Models for Medical Imaging Classification

The few-shot metric learning models have been evaluated on the Meta-Dataset. With the introduction of fsMIB, it is important to analyse whether these models continue to achieve relevant results and whether their historical performance trends are maintained. In an effort to facilitate this comparison, the Average Rank of each model was plotted as a function of the number of shots in both the **meta-med** and **med-med** experiments. These are presented in Figure 6.

From these plots, it is possible to conclude that TSA is the current state-of-the-art model for few-shot metric learning on the fsMIB, with the best average rank across the number of shots in both experiments, except for 1 and 5 shots in the **med-med** experiment, where its average rank is equal to or worse than that of SCNAPS. Nonetheless, after a thorough analysis of Table A.2, in the Appendix of the Master's thesis [45], it is clear that, in these exception cases, TSA achieves higher accuracies for a higher number of datasets compared to SCNAPS. However, its rank in the **shoulder** dataset is lower, which negatively impacts its average rank. All things considered, these results support the findings from the Meta-Dataset reports, where TSA was also shown to be the top performing model.

It was expected that ProtoNet would perform worse than the other models, owing to being one of the first metric learning classification models in FSL. All the same, in the **med-med** experiment there are two exceptions where URL and TCNAPS performed worse than ProtoNet.

According to Section VII-B, URL is performing poorly in the **med-med** experiment, despite achieving the second place in the Meta-Dataset evaluation. The results shown in Figure 6 indicate that the performance of URL is close to random guessing, unlike the performance of the other models. The exact cause of these results is not entirely understood. Still, it is believed to be related to either a poor choice of hyperparameters or an inherent problem with knowledge distillation. One potential contributing factor could be the lack of an extensive hyperparameter tuning for the training of the domain-specific networks. Another possible cause could be the better generalization achieved by using the Meta-Dataset for training, compared to using the fsMIB. Training on the fsMIB has been shown to struggle with generalization to other medical image classification problems, as highlighted in Section VII-C. As a result, the domain-specific networks

may have difficulty generalizing, since they are only trained on their specific domain images. This may prevent the distillation of generalizable features to the task-agnostic feature extractor. It is noteworthy that, in spite of being based on the feature extractor of URL which is performing poorly, TSA still performs as the state-of-the-art model in the **med-med** experiment. This suggests that adapters, especially the residual adapters used in TSA, provide an effective method for adapting the feature extractor to new tasks, albeit at the cost of additional training and inference time. One example of this is visible by comparing URL and TSA plots of Figure 4.

Historically, when applied to Meta-Dataset, SCNAPS, TCNAPS and URL have been the state-of-the-art models, in this order. In this article, they were evaluated on the fsMIB. Contrary to expectations, in the **meta-med** experiment, they achieved equivalent average ranks and, in the **med-med** experiment, URL performed poorly, while SCNAPS outperformed TCNAPS for 1, 5 and 15 shots. The poor performance of URL in the **med-med** experiment was already discussed. The reason of the similar mean accuracies achieved by SCNAPS and TCNAPS compared to URL in the **meta-med** experiment may be related to the knowledge distillation process of URL being more effective when there is a larger similarity between the training and evaluation domains. Finally, the reason for SCNAPS achieving better results than TCNAPS in the lowest shot scenario is likely due to two factors. First, the benefits of TCNAPS are greater when there are more images in the query set. However, the number of images per class in the query set is fixed. Second, the algorithm used by TCNAPS to update the Mahalanobis distance parameters has a higher probability of working better with a larger amount of images in the support set. As the number of shots increases, the update algorithm will have more labeled information a priori, which will improve the class predictions used in the query set and lead to improved updates. This can be seen in Figure 6 for the **med-med**, where TCNAPS's average rank improves with the increase in the number of shots, while the opposite happens to SCNAPS.

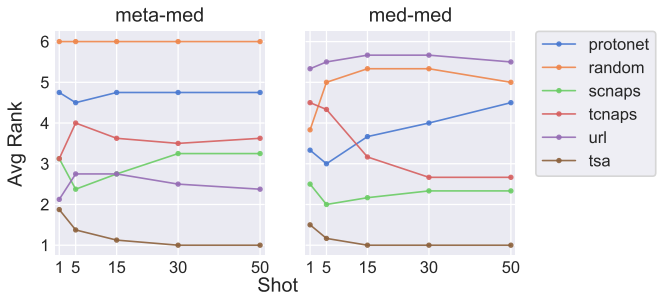


Fig. 7: Plots of the Average Rank, given the number of shots, for each of the few-shot image classification models and for each of the experiments, without including the evaluation of the MURA datasets.

Considering that some MURA datasets are difficult, given the training and testing conditions stated in Section VII, it is important to improve the previous analysis without taking into consideration the noise introduced by these datasets. The average ranks for each classification model and for each experiment are displayed in Figure 7. In the **meta-med** experiment, all models performed better than random guessing for all shots, resulting in the random baseline always having an average rank of 6. In the **med-med** experiment, URL continued to perform poorly. Additionally, for both the experiments, SCNAPS always achieved a smaller average rank than the one from TCNAPS, contradicting the literature [15], [16]. Nevertheless, it is worth considering that TCNAPS advantages were not fully explored.

E. Evaluation of Few-Shot Image Classification Models: Limitations and Suggestions for Improvement

FSL has seen rapid development in recent years, as it represents an area where machines have difficulties to outperform humans. The literature [6], [7], [20] has been reporting results using the mean accuracy over 600 test episodes and the corresponding 95% confidence intervals as evaluation metrics. However, this metrics may not provide the necessary information to evaluate and assess the FSL models suitability for a particular application, for example, medical imaging diagnostics. The mean accuracy does not reveal the variation in accuracy among the 600 test episodes and the 95% confidence intervals may not be intuitive enough to understand the amount of variation present in the classification task. Therefore, the standard deviation, which allows humans to quickly understand the dispersion of the accuracies, would be a more intuitive metric.

The minimum value of the mean standard deviation, obtained across the evaluation datasets for each shot and model in the **meta-med** and **med-med** experiments, was 7.1%, indicating a significant amount of variation in the results. Examining a random case, specifically, the results obtained by TSA in the 600 test episodes, with 50 shots, randomly sampled from **shoulder** dataset, represented in Figure 8, it is clear that there is a wide range of accuracies ranging from a minimum of 30% to a maximum of 83.33%, with an average of 56.42% and a standard deviation of 9.03%. This standard deviation corresponds to a 95% confidence interval of 0.72 %. In the literature, this result would be reported as 56.42 ± 0.72 %, hiding important information needed to evaluate the performance of few-shot image classification models. This suggests that to effectively compare these models, it would be beneficial to report results using additional metrics, such as the five-number summary – minimum, first quartile, median, third quartile and maximum –, as a complement to the mean accuracy and confidence intervals.

Besides the problems related to the evaluation metrics, there is a problem with the current approach to evaluate few-shot image classification methods, which involves randomly sampling episodes. This approach leads to a wide spread of accuracies, owing to the inherent variation in the characteristics of the sampled episodes. This variation can impact the adequacy of the support set for classifying a query set, as a result of the variability of the images within a dataset. It is challenging to accurately evaluate the effectiveness and suitability of a FSL method for use in medical imaging

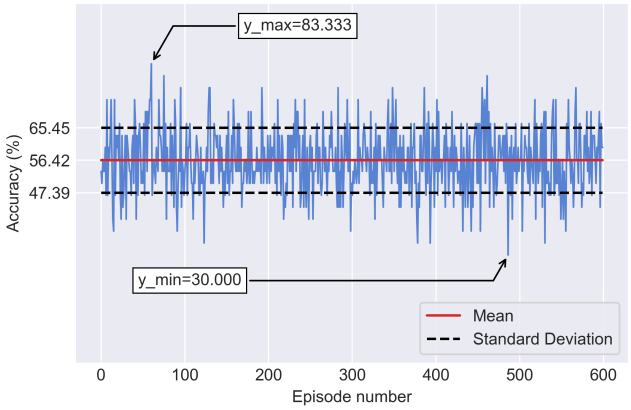


Fig. 8: Accuracies obtained by TSA in the 600 test episodes, with 50 shots, randomly sampled from **shoulder** dataset.

diagnostics when it is evaluated using 600 test episodes with different and random compositions. To address this issue, alternative evaluation methods, such as fixing the support set and evaluating all other images as queries, may provide a more robust and realistic evaluation method by approximating a real-world scenario, where the support set is not randomly chosen, but rather carefully selected by a user. Moreover, examining the characteristics of a good support set and how it can affect the performance of a few-shot learning method can provide valuable insights to improve the use of these methods in medical imaging classification tasks. For instance, a support set that is representative of the full distribution of the dataset may be more effective than a support set that is not representative.

All in all, the use of mean accuracy and confidence intervals as evaluation metrics, as well as the current approach to evaluate few-shot image classification models, may not provide a comprehensive understanding of the performance of these models in medical imaging classification tasks. As a consequence, FSL requires further research in these areas.

VIII. Conclusion

In this work, the fsMIB is proposed as a valuable tool for the study of few-shot image classification in the medical imaging field, as a result of its public availability and ease of use, as well as the inclusion of a variety of visual concepts – radiological, ultrasound and dermatoscopic images.

In general, the few-shot metric learning models performed better than the random baseline. Nevertheless, these models struggled with classifying MURA datasets, achieving accuracies similar to random guessing. The experiments confirmed that as the number of shots increases, the accuracy of the models improves, but the rate of accuracy improvement decreases. Another conclusion drawn was that training on the meta-dataset benchmark lead to higher accuracies than training on the fsMIB.

It was found that while the mean accuracy over 600 test episodes and the corresponding 95% confidence intervals can provide some insight into the performance of the models, they may not capture all the important information needed to fully evaluate them. For example, the mean accuracy can be misleading, on account of the wide spread of accuracies observed in the results. In this case, using additional metrics, such as the five-number summary, would provide a more comprehensive understanding of the model's performance. Also, the current approach to evaluate few-shot image classification models causes, inherently, a wide spread of accuracies in the results, making it challenging to accurately assess the effectiveness and suitability of these models for use in medical imaging diagnostics. This is particularly concerning in the medical imaging field, where accurate diagnoses are critical for patient care.

IX. Future Work

FSL is an active research area in DL. Whilst conducting this research, several future opportunities emerged:

- Improve the benchmark by finding more effective preprocessing techniques for each dataset;
- Evaluate the models on fsMIB using the Meta-Dataset sampling algorithm to enable a more accurate comparison with the Meta-Dataset results;
- Conduct experiments using data augmentation during training, as in the URL and TSA models, and compare the FSL models evaluation results;
- Improve the composition of support sets to avoid underrepresentation and deficiencies of training data [52], by looking for the key characteristics of a support set and how these characteristics impact performance;
- Conduct experiments using multiple combinations of train and test fsMIB datasets and compare the FSL models evaluation results;
- Conduct experiments using traditional transfer learning models to train and evaluate on the fsMIB datasets and compare these models with the FSL models.
- Increase the number of shots in the support set of each episode through data augmentation to allow the models to perform better, as the model's performance tends to improve with an increase in the number of shots;
- Discover ways to explain FSL models predictions. For example, by applying GradCAM [53]. Without explainable and interpretable methods, clinicians may be hesitant to adopt these methods in the medical imaging analysis context [52].
- Explore the connection between self and semi-supervised learning and few-shot learning, since TCNAPS provides a perspective on using the unlabeled instances.
- Conduct experiments of pretraining the models on the Meta-Dataset benchmark and, then, training and evaluating them on the fsMIB, since it was found that under the training and testing details, described in Section VII, the models were able to generalize better when trained on the Meta-Dataset.

REFERENCES

[1] G. Leo, "Challenges in estimating reproducibility of imaging modalities." *World journal of methodology*, vol. 1 1, pp. 12–4, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4145555/

[2] A.-R. A. Ali and T. M. Deserno, "A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data," in *Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, C. K. Abbey and C. R. Mello-Thoms, Eds., vol. 8318, Feb. 2012, p. 83181I. [Online]. Available: https://doi.org/10.1117/12.912389

[3] H. Chang, "Skin cancer reorganization and classification with deep neural network," *CoRR*, vol. abs/1703.00534, 2017. [Online]. Available: http://arxiv.org/abs/1703.00534

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[5] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," *CoRR*, vol. abs/1906.02243, 2019. [Online]. Available: http://arxiv.org/abs/1906.02243

[6] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 3630–3638. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html

[7] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *CoRR*, vol. abs/1703.05175, 2017. [Online]. Available: http://arxiv.org/abs/1703.05175

[8] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=rJY0-Kcll

[9] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 1126–1135. [Online]. Available: http://proceedings.mlr.press/v70/finn17a.html

[10] N. Dvornik, C. Schmid, and J. Mairal, "Selecting relevant features from a universal representation for few-shot classification," *CoRR*, vol. abs/2003.09338, 2020. [Online]. Available: https://arxiv.org/abs/2003.09338

[11] L. Liu, W. L. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=04cII6MumYV

[12] J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner, "Fast and flexible multi-task classification using conditional neural adaptive processes," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 7957–7968. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/1138d90ef0a0848a542e57d1595f58ea-Abstract.html

[13] Y. Liu, J. Lee, L. Zhu, L. Chen, H. Shi, and Y. Yang, "A multi-mode modulator for multi-domain few-shot classification," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 8433–8442. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.00834

[14] E. Triantafillou, H. Larochelle, R. S. Zemel, and V. Dumoulin, "Learning a universal template for few-shot dataset generalization," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 10 424–10 433. [Online]. Available: http://proceedings.mlr.press/v139/triantafillou21a.html

[15] P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal, "Improved few-shot visual classification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 14 481–14 490. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Bateni_Improved_Few-Shot_Visual_Classification_CVPR_2020_paper.html

[16] P. Bateni, J. Barber, J. van de Meent, and F. Wood, "Enhancing few-shot image classification with unlabelled examples," in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*. IEEE, 2022, pp. 1597–1606. [Online]. Available: https://doi.org/10.1109/WACV51458.2022.00166

[17] W. Li, X. Liu, and H. Bilen, "Universal representation learning from multiple domains for few-shot classification," *CoRR*, vol. abs/2103.13841, 2021. [Online]. Available: https://arxiv.org/abs/2103.13841

[18] W.-H. Li, X. Liu, and H. Bilen, "Cross-domain Few-shot Learning with Task-specific Adapters," *arXiv e-prints*, p. arXiv:2107.00358, Jul. 2021. [Online]. Available: https://arxiv.org/abs/2107.00358

[19] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, pp. 1332 – 1338, 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26659050/

[20] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P. Manzagol, and H. Larochelle, "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=rkgAGAVKPr

[21] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, pp. 611 – 629, 2018. [Online]. Available: https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9

[22] F. Shakeri, M. Boudiaf, S. Mohammadi, I. Sheth, M. Havaei, I. Ben Ayed, and S. Ebrahimi Kahou, "FHIST: A Benchmark for Few-shot Classification of Histological Images," *arXiv e-prints*, p. arXiv:2206.00092, May 2022. [Online]. Available: https://arxiv.org/abs/2206.00092

[23] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 1199–1208. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Sung_Learning_to_Compare_CVPR_2018_paper.html

[24] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, Eds. ACM, 2006, pp. 535–541. [Online]. Available: https://doi.org/10.1145/1150402.1150464

[25] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[26] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 3942–3951. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16528

[27] S. Rebuffi, H. Bilen, and A. Vedaldi, "Efficient parametrization of multi-domain deep neural networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 8119–8127. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Rebuffi_Efficient_Parametrization_of_CVPR_2018_paper.html

[28] Y. Guo, N. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris, "A broader study of cross-domain few-shot learning," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part*

XXVII, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12372. Springer, 2020, pp. 124–141. [Online]. Available: https://doi.org/10.1007/978-3-030-58583-9_8

[29] R. Singh, V. Bharti, V. Purohit, A. Kumar, A. K. Singh, and S. K. Singh, "Metamed: Few-shot medical image classification using gradient-based meta-learning," *Pattern Recognition*, vol. 120, p. 108111, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321002983

[30] M. Shorfuzzaman and M. S. Hossain, "Metacovid: A siamese neural network framework with contrastive loss for n-shot diagnosis of covid-19 patients," *Pattern Recognition*, vol. 113, p. 107700, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320305033

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

[32] S. Maji, E. Rahtu, J. Kannala, M. B. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *CoRR*, vol. abs/1306.5151, 2013. [Online]. Available: http://arxiv.org/abs/1306.5151

[33] C. Wah, S. Branson, P. Welinder, P. Perona, and S. J. Belongie, "The caltech-ucsd birds-200-2011 dataset," in *Computation & Neural Systems Technical Report*, 2011. [Online]. Available: http://authors.library.caltech.edu/27452/1/CUB_200_2011.pdf

[34] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," *CoRR*, vol. abs/1311.3618, 2013. [Online]. Available: http://arxiv.org/abs/1311.3618

[35] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, and N. Fox-Gieg. (2016) The Quick, Draw! – A.I. experiment. [Online]. Available: https://quickdraw.withgoogle.com

[36] B. Schroeder and Y. Cui. (2018) FGVCx fungi classification challenge 2018. [Online]. Available: https://github.com/visipedia/fgvcx_fungi_comp

[37] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008, pp. 722–729. [Online]. Available: https://ieeexplore.ieee.org/document/4756141

[38] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The german traffic sign detection benchmark," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/6706807

[39] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[40] W. Chen, Y. Liu, Z. Kira, Y. F. Wang, and J. Huang, "A closer look at few-shot classification," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=HkxLXnAcFQ

[41] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=rylXBkrYDS

[42] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 9042–9051. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.00893

[43] S. Laenen and L. Bertinetto, "On episodes, prototypical networks, and few-shot learning," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 24 581–24 592. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/cdfa4c42f465a5a66871587c69fcfa34-Abstract.html

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[45] L. Tourais, "Medical imaging benchmark for few-shot image classification," 2023.

[46] A. M. Tahir, M. E. H. Chowdhury, Y. Qiblawey, A. Khandakar, T. Rahman, S. Kiranyaz, U. Khurshid, N. Ibtehaz, S. Mahmud, and M. Ezeddin, "Covid-qu-ex ." 2021. [Online]. Available: https://doi.org/10.34740/kaggle/dsv/3122958

[47] W. Al-Dhabyani, M. M. Gomaa, H. Khaled, and A. A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31867417/

[48] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs," *arXiv e-prints*, p. arXiv:1712.06957, Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.06957

[49] S. Jaeger, S. Candemir, S. K. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. R. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases." *Quantitative imaging in medicine and surgery*, vol. 4 6, pp. 475–7, 2014. [Online]. Available: https://lhncbc.nlm.nih.gov/publication/pub9356

[50] N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. A. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *CoRR*, vol. abs/1902.03368, 2019. [Online]. Available: http://arxiv.org/abs/1902.03368

[51] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *CoRR*, vol. abs/1803.10417, 2018. [Online]. Available: http://arxiv.org/abs/1803.10417

[52] V. Rotemberg, A. Halpern, S. Dusza, and N. C. Codella, "The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice." in *Seminars in Cutaneous Medicine and Surgery*, vol. 38, no. 1, 2019, pp. E38–E42.

[53] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391