Navigating Data Scarcity using Foundation Models: A Benchmark of Few-Shot and Zero-Shot Learning Approaches in Medical Imaging

Stefano Woerner[1] and Christian F. Baumgartner[1,2]

[1] Cluster of Excellence "Machine Learning", University of Tübingen, Germany
{firstname.lastname}@uni-tuebingen.de
[2] Faculty of Health Sciences and Medicine, University of Lucerne, Switzerland

Abstract. Data scarcity is a major limiting factor for applying modern machine learning techniques to clinical tasks. Although sufficient data exists for some well-studied medical tasks, there remains a long tail of clinically relevant tasks with poor data availability. Recently, numerous foundation models have demonstrated high suitability for few-shot learning (FSL) and zero-shot learning (ZSL), potentially making them more accessible to practitioners. However, it remains unclear which foundation model performs best on FSL medical image analysis tasks and what the optimal methods are for learning from limited data. We conducted a comprehensive benchmark study of ZSL and FSL using 16 pretrained foundation models on 19 diverse medical imaging datasets. Our results indicate that BiomedCLIP, a model pretrained exclusively on medical data, performs best on average for very small training set sizes, while very large CLIP models pretrained on LAION-2B perform best with slightly more training samples. However, simply fine-tuning a ResNet-18 pretrained on ImageNet performs similarly with more than five training examples per class. Our findings also highlight the need for further research on foundation models specifically tailored for medical applications and the collection of more datasets to train these models.

1 Introduction

Machine learning is revolutionizing the field of medical imaging and diagnostics, offering capabilities that were previously unattainable. However, these advancements typically depend on the availability of large, well-annotated datasets. For many medical applications, such as the diagnosis of rare diseases, collecting these types of datasets is often infeasible. Consequently, in many real-world scenarios, there is often insufficient data to effectively train highly performant deep learning models. Additionally, computational resources are frequently limited, which poses further challenges in training or even fine-tuning state-of-the-art models.

Few-shot learning (FSL) has shown great potential in addressing these data-scarce applications. With effective FSL strategies, clinics and medical researchers could potentially train models using their own small datasets and achieve performance levels acceptable for clinical practice. Few-shot learning is most commonly

performed through fine-tuning of a large pretrained model on the smaller, domain-specific, target dataset. Recently, several large models, known as foundation models, have been published after being trained on vast amounts of data [9, 2, 8, 16]. Many such models have been shown to have excellent generalization capabilities, and to be highly suitable for FSL. However, no large-scale studies exist which compare FSL performance of different pretrained models across a broad and diverse array of medical imaging domains. A number of foundation models are also capable of zero-shot learning (ZSL) by searching for the highest correspondence between the representations of the input image and a language prompt. Similarly, there are no works rigorously comparing the ZSL capabilities of different foundation models on a diverse range of medical tasks.

In this paper, we present the first large-scale study comparing the FSL and ZSL performance of various publicly available pretrained models across a diverse set of medical imaging domains. We conduct our study on the recently released MedIMeta dataset [15], which is comprised of 19 different datasets from 10 different imaging modalities and anatomical regions. In comprehensive experiments we evaluate 16 publicly available models that have been pretrained on different medical and non-medical data sources. Because fine-tuning very large models is not practical within the computational budget of most clinicians and researchers, we limited ourselves to exploring strategies that are possible to perform in the realistic scenario of having access to a single mid-range to high-end GPU. Within these constraints we explore a linear probing strategy as well as fine-tuning. For the five models in our benchmark that support ZSL, we also benchmark their ZSL capabilities with different prompt styles.

Our experiments yield a number of practical insights and actionable recommendations. We make code to reproduce our results and adapt our experiments publicly available.[3]

2 Methods

2.1 Dataset

To allow us to study the FSL and ZSL performance on wide array of different image modalities and tasks, we conduct our experiments on the recently released MedIMeta dataset [15, 14]. MedIMeta is a highly standardized meta-dataset compiled from 19 publicly available datasets, and covering 10 different imaging modalities. We use the main (i.e. first) task for each of the 19 datasets. We refer the reader to [15, 14] for detailed descriptions of the sub-datasets and tasks.

2.2 Simulation of FSL and ZSL tasks

We artificially construct multiple FSL tasks from each of MedIMeta's datasets by randomly sampling n labeled training samples and 10 unlabeled query samples per class from each dataset. We ensure that no images of the same subject are

[3] `https://github.com/StefanoWoerner/medimeta-fsl-benchmark`

spread over the two sets. We coin these individual FSL tasks a *task instance*. To ensure robust FSL performance measurement, we randomly generate 100 task instances for each dataset and average the results. In order to investigate the effect of increasing numbers of labeled training samples we repeat all experiments for $n \in \{1, 2, 3, 5, 7, 10, 15, 20, 25, 30\}$. In addition we also simulate task instances with $n = 0$, i.e. only query samples, for the ZSL evaluation.

2.3 Pretrained Models

We evaluate three distinct pretraining paradigms: supervised pretraining, self-supervised pretraining, and contrastive language-image pretraining (CLIP). In the following we briefly describe the specific architectures and pretraining data.

Fully Supervised Models. We investigate the widely used **Residual Networks (ResNet) architecture** [6] in the variations ResNet18, ResNet50, and Resnet101, all of which have been pretrained on the ImageNet dataset [11].

We further investigate the **Vision Transformer (ViT) architecture** [4]. Due to it's excellent performance on many computer vision benchmarks, the ViT has become a standard architecture and the basis of a large amount of further work. We compare the base (ViT-B), large (ViT-L), and huge (ViT-H) architecture variations with patch sizes 16, 16 and 14, respectively. We consider models pretrained on ImageNet [11] and on ImageNet21k [10].

Self-supervised Models. In this category we consider the **self-DIstillation with NO labels (DINO)** model [1]. We specifically focus on the recently released DINOv2 model [8] which relies on a ViT architecture that was pretrained using a self-supervised knowledge distillation approach. The model was trained using a very large unlabeled but curated dataset assembled from various computer vision datasets. The DINOv2 representations have been shown to be highly transferable across computer vision tasks [8]. We consider the ViT-B, ViT-L, and giant (ViT-g) variations with patch size 14.

Contrastive Language-Image Pretraining. Lastly, we consider two CLIP models which employ contrastive learning to align images and text into a shared embedding space [9].

Firstly, we use the original **CLIP model** with the weights for ViT-B and ViT-L provided by OpenAI [9]. These models have been pretrained on 400 million image-text pairs collected from the internet. Although the specific composition of this dataset is not traceable, it is likely that a small portion of medical data was included. In addition to its unique ZSL capabilities, the CLIP model was also shown to perform extremely well on computer vision FSL tasks by training a linear probe on the final image-encoder representations [9].

Secondly, we use the ViT-H and ViT-g models trained on LAION-2B [12] provided by OpenCLIP [2], an open source reimplementation of OpenAI's CLIP. LAION-2B contains 2 billion image-text pairs extracted from common crawl [3] and is the English language subset of the larger LAION-5B [12] dataset. Similar to the OpenAI data, the inclusion of small amounts of medical data is likely.

We also investigate the **BiomedCLIP model** [16] which uses the same ViT architecture as the base version of the original CLIP, but replaces the text encoder with PubMedBERT [5], a language model tailored for the biomedical domain. BiomedCLIP was pretrained on 15 million text-image pairs extracted from PubMed articles (PMC-15M). This is the only model in our study that was trained exclusively on medical data. BiomedCLIP can be employed for FSL and prompt-based ZSL in the same manner as CLIP.

2.4 Few-shot Learning Strategies

We evaluate two model adaptation strategies: fine-tuning and linear probing.

Fine-tuning involves initializing a network with pretrained weights, and then continuing the training of all weights in the network with an FSL task instance. The last linear layer (classification layer) is replaced with a new layer matching the number of classes in the target task. For most foundation models, which commonly have hundreds of millions or even billions of parameters, fine-tuning is computationally infeasible for many practitioners. We therefore only evaluate the fine-tuning strategy on the ResNet-18 and ResNet-50 variants.

Similarly, **linear probing** involves initializing a network with pretrained weights, and attaching a new classification layer. However, in linear probing the backbone network is frozen, and a simple linear classifier is trained on the final representations of the network. This was shown to lead to strong FSL performance assuming the base network is able to extract useful image features [9]. Since only a linear classifier is trained on the image features produced by the pretrained network, this strategy is computationally much cheaper than fine-tuning the complete network, making it feasible to use with large foundation models.

We conduct an extensive **hyper-parameter search** on a separate set of sampled FSL tasks for both fine-tuning and linear probing. For each of the models and each number of labeled samples n, we test two optimizers (SGD and Adam), two different head initialization strategies (Kaiming initialization [6], initialization with all zeros [13]), a range of learning rates between 10^{-5} and 0.1, and a range of training steps between 5 and 200. We evaluate all models from Section 2.3 using their respective optimal parameters from the hyper-parameter search.

2.5 Zero-shot learning strategy

The CLIP [9] and BiomedCLIP [16] models have the capability of solving classification tasks with no labeled training examples by searching for the highest similarity between an input image and several text prompts corresponding to different target classes. We test three different prompt templates. First, we investigate simply using the class names extracted from the MedIMeta task definitions as prompts. Secondly, we test two templates which add information about the imaging modality: "A {domain_identifier} image where the {task_name} is {class_name}", and "This {domain_identifier} image shows [a] {class_name}".

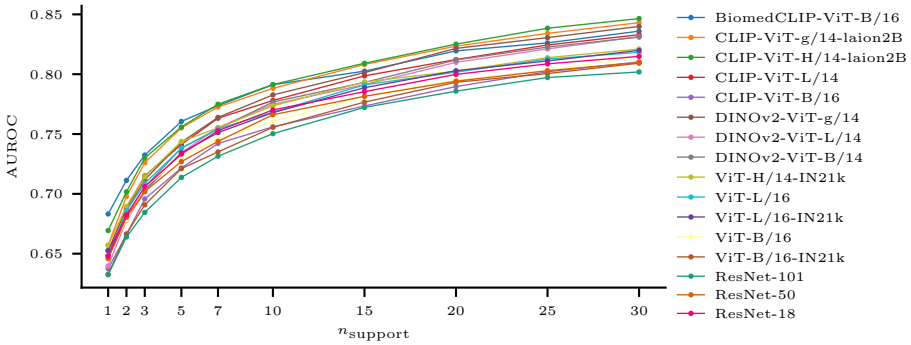

Fig. 1. Harmonic mean AUROC over all 19 MedIMeta datasets

All variables above are extracted from the MedIMeta task description files. However, some class names and domain identifiers needed to be adjusted in order to form a grammatically correct and semantically meaningful sentences.

2.6 Metrics

We evaluate the performance for each dataset and each training set size n using the area under the receiver operator curve (AUROC) averaged over all 100 task instances. To obtain a measure of average performance across all datasets, we use the harmonic mean of the AUROCs from each dataset.

3 Experiments and Results

We performed all FSL and ZSL experiments using all models and learning strategies as described above. In the following, we describe our main findings. All results can be found in Table A.1 in the Supplementary Material.

The optimal hyperparameters were similar for all models. For all models the best-performing optimizer was Adam [7]. Further, initializing the classification head with zeros performed better or on par compared to Kaiming initialization [6], in line with the findings in [13]. For most models and n, using a learning rate of 10^{-4} with at least 120 training steps was optimal or close to optimal.

Linear probing with BiomedCLIP and CLIP-ViT-H yielded the best results on average. As can be seen in Fig. 1 CLIP-ViT-H on average outperformed other pretrained models for $n \geq 7$. Interestingly, the performance of the "huge" variant of CLIP was slightly better than the even larger "giant (g)" variant. However, for smaller n, BiomedCLIP, which was the only foundation model in our comparison trained entirely with medical data, outperformed its larger CLIP counterparts and performed on par with CLIP-ViT-H up to $n = 10$. We note that CLIP-based pretraining led to the best performance overall, underscoring the potential of contrastive language-image pretraining for FSL.

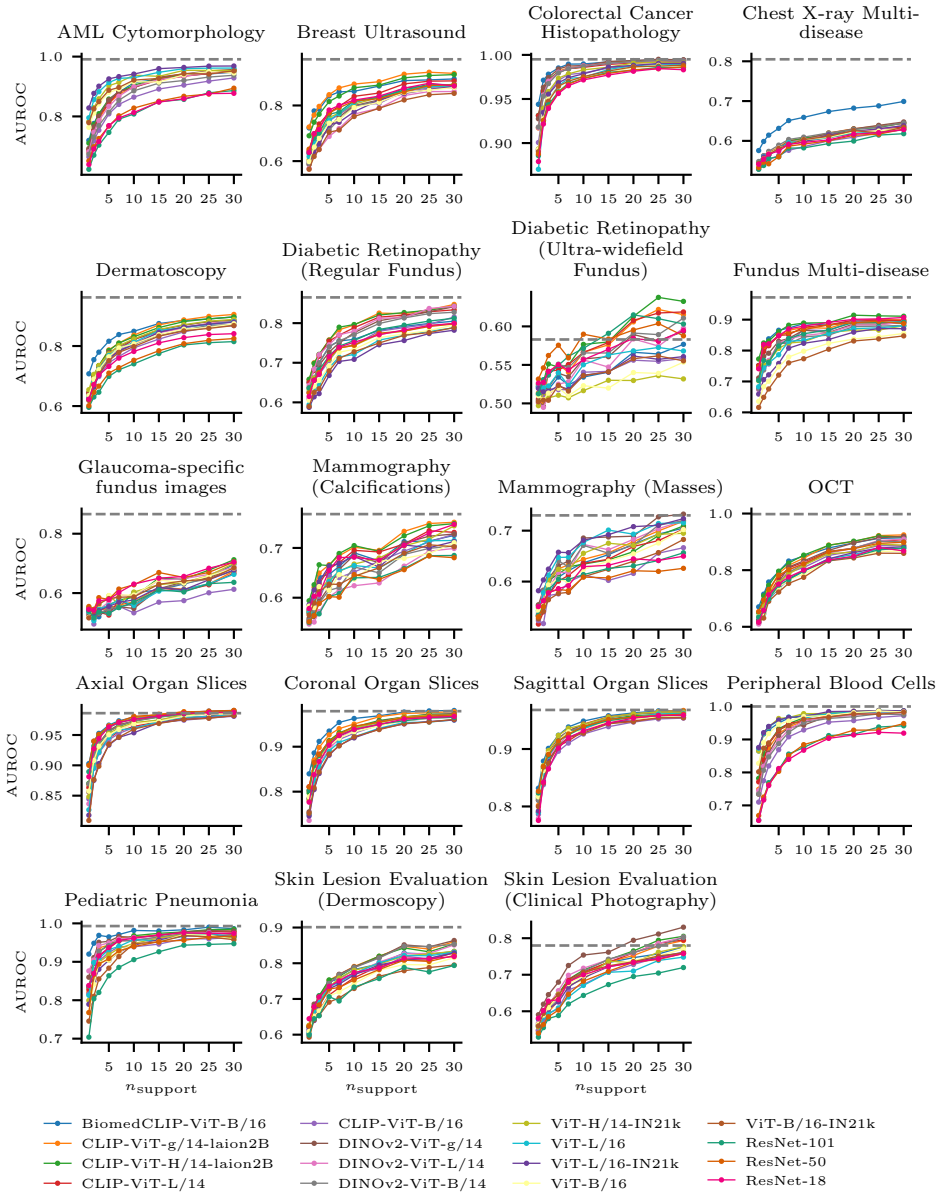

Fig. 2. AUROC on the different datasets with fully supervised baseline from [15]. The fully-supervised performance is indicated by the black dotted line.

Linear probing performance on individual datasets was mixed. The performance on the individual datasets shown in Fig. 2 was mixed. Interestingly, the method that performed best on average for $n \geq 7$ (CLIP-ViT-H) rarely

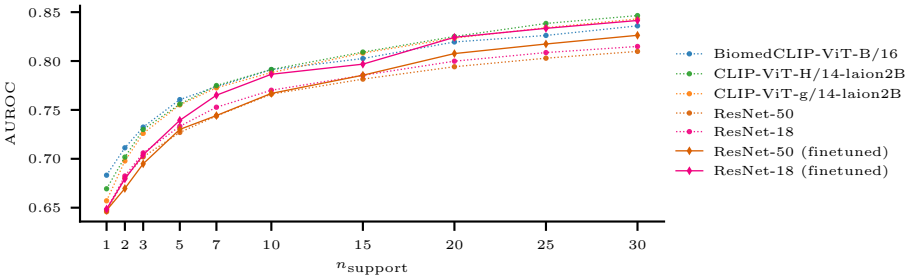

Fig. 3. Harmonic mean AUROC across all 19 MedIMeta datasets of fine-tuned ResNet models with the best-performing linear probe as a point of comparison.

performed the best on the individual datasets. Rather it was consistently among the top-few approaches on most datasets leading to its high average performance. BiomedCLIP, on the other hand, performed very well on some datasets (e.g. Chest X-ray Multi-disease, Dermatoscopy, and Pediatric Pneumonia), but more poorly on others (e.g. Ultra-widefield Fundus, or Mammography (Masses)). We hypothesize that BiomedCLIP performed more strongly on images that were overrepresented in the PMC-15M pretraining dataset. We conclude that in practice linear probing on the BiomedCLIP model might often be a good first attempt when working with very few labeled images, but it does not obviate thorough evaluation on a held-out test set. With more training data, linear probing CLIP-ViT-H is likely the better option, especially since it does not display as much variablility throughout different imaging modalities as BiomedCLIP.

FSL performed close to fully supervised learning for some tasks. In Fig. 2 we additionally show the fully supervised baseline performance on the official data splits reported in [14]. We observed that for some tasks the 30-shot performance almost matched the fully supervised performance. Indeed, for the Ultra-widefield Fundus dataset the FSL performance was substantially better than the fully supervised performance. We believe this is due to the small number of training images in the official split of this dataset. Nevertheless, this suggests that for some problems linear probing of a foundation model may be a better alternative than training a model from scratch with a small dataset.

Linear probes on large models beat fine-tuning of small models. In Fig. 3 it can be seen that linear probing with CLIP-ViT-H on average outperformed fine-tuning of the ResNet-18 and ResNet-50 for all n. However, with more training data fine-tuning the ResNet-18 performed *almost* as well as CLIP-ViT-H, and for $n \geq 20$ the fine-tuned ResNet-18 outperformed BiomedCLIP. Interestingly, fine-tuning ResNet-18 clearly outperformed fine-tuning ResNet-50, suggesting that a lower network complexity may be preferable in the FSL scenario. Our findings suggest that while linear probing very large foundation models such as the CLIP-ViT-H on average may lead to small performance gains, the commonly used strategy of fine-tuning a ResNet-18 also performs strongly given sufficient

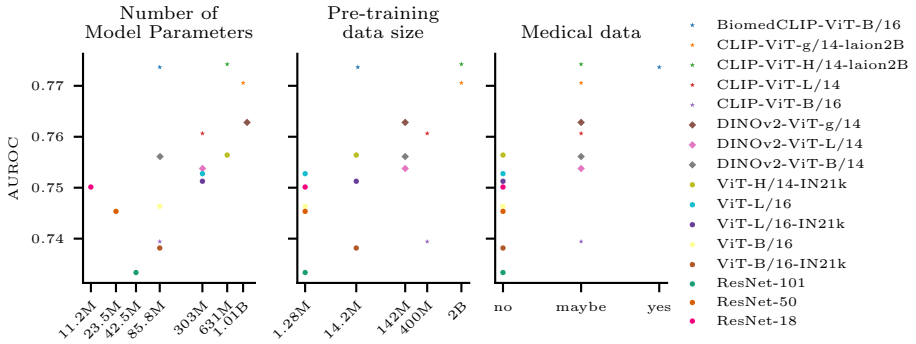

Fig. 4. Model properties plotted against mean few-shot performance over the tested n. The symbol indicates the type of pretraining (Supervised, DINO, CLIP).

data. We note that while fine-tuning of foundation models may lead to even better results, this is computationally prohibitive for the majority of practitioners.

ZSL performance could not match FSL performance. The ZSL approaches had average AUROC scores ranging from 0.316 to 0.397, far below those of the 1-shot performance reported in Fig. 1 where average AUROCs range from 0.632 to 0.683. This contradicts the findings of Radford et al. [9] who showed that on computer vision tasks the CLIP model can often outperform linear probes in the ZSL setting. We conclude that ZSL may not yet be a suitable strategy for general medical image analysis tasks. We report the ZSL results in Fig. A.1 in the Supplementary Materials.

Model complexity and pretraining data size correlate with performance. In Fig. 4, we explore the relation of the following three model properties to their linear probing performance: the model size, the number of samples in the pretraining data, and the type of pretraining data. We observed that there was a strong positive correlation between model size and few-shot performance as well as pretraining set size and few-shot performance. While the non-medical CLIP-ViT-H, and CLIP-ViT-g clearly outperformed all other non-medical approaches on average, BiomedCLIP, which was trained on medical data exclusively, performed very well despite much smaller number of parameters and pretraining data size. This underscores the need for building training sets which contain a diverse set of medical images and for training advanced medicine-focused foundation models.

4 Conclusion

We performed the first large-scale study comparing the FSL and ZSL performance of a wide array of pretrained models on a diverse set of medical imaging data. We found that, on average, in the very low data regime of $n \leq 5$ samples per class, a linear probe on BiomedCLIP was the best strategy. However, with more data, linear probing of the CLIP-ViT-H model performed slightly better. While

fine-tuning a ResNet-18 on average performed worse compared to a linear probe on CLIP-ViT-H, it still reached a high performance for $n \geq 20$. We also observed a large variance between the performance on the different datasets emphasizing the need for cautious application of these technologies. Our investigation further revealed that parameter-rich foundation models trained on very large non-medical datasets have very good FSL performance on medical tasks. However, the strong performance of BiomedCLIP model on some datasets underscores the potential of foundation models specific to medical applications.

Acknowledgments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Stefano Woerner.

Disclosure of Interests. The authors have no competing interests to declare.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. "Reproducible Scaling Laws for Contrastive Language-Image Learning". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. DOI: 10.1109/cvpr52729.2023.00276. URL: http://dx.doi.org/10.1109/CVPR52729.2023.00276.

[3] *Common Crawl*. https://commoncrawl.org.

[4] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020. arXiv: 2010.11929 [cs.CV].

[5] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ACM Transactions on Computing for Healthcare* 3.1 (Oct. 2021), pp. 1–23. ISSN: 2637-8051. DOI: 10.1145/3458754. URL: http://dx.doi.org/10.1145/3458754.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[7] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[8] Maxime Oquab et al. "Dinov2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023).

[9] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[10] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. *ImageNet-21K Pretraining for the Masses*. 2021. arXiv: 2104.10972 [cs.CV].

[11] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[12] Christoph Schuhmann et al. *LAION-5B: An open large-scale dataset for training next generation image-text models*. 2022. arXiv: 2210.08402 [cs.CV].

[13] Stefano Woerner and Christian F. Baumgartner. "Strategies for Meta-Learning with Diverse Tasks". In: *Medical Imaging with Deep Learning*. 2022.

[14] Stefano Woerner, Arthur Jaques, and Christian Baumgartner. *MedIMeta: A comprehensive and easy-to-use multi- domain multi-task medical imaging meta-dataset*. Zenodo, Apr. 2024. DOI: 10.5281/zenodo.7884735. URL: https://doi.org/10.5281/zenodo.7884735.

[15] Stefano Woerner, Arthur Jaques, and Christian F. Baumgartner. *A comprehensive and easy-to-use multi-domain multi-task medical imaging meta-dataset (MedIMeta)*. 2024. arXiv: 2404.16000 [cs.CV].

[16] Sheng Zhang et al. *BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs*. 2023. arXiv: 2303.00915 [cs.CV].