MoPET: Parameter-Efficient Mixture-of-Experts for Unified Medical Image Classification

Sebastian Doerrich*, Daniel Würtinger*, Francesco Di Salvo, Shyam Nandan Rai, and Christian Ledig

xAILab Bamberg, University of Bamberg, Bamberg, Germany
sebastian.doerrich@uni-bamberg.de

Abstract. Adapting deep learning models to profound clinical heterogeneity typically relies on parameter-efficient fine-tuning (PEFT) to avoid the severe overfitting associated with full end-to-end network updates. Although PEFT successfully navigates limited data scenarios, it inherently forces the training of a separate, isolated adapter for every specific diagnostic task. Consolidating these isolated adapters into a single generalist network risks negative transfer, as optimization gradients from conflicting visual domains interfere. To address this, we propose *MoPET*, a mixture-of-experts (MoE) method that uses a learned sparse router to direct each input through a small subset of low-rank PEFT experts injected into a frozen foundation model, sharing capacity across datasets while limiting cross-domain gradient conflict. Through selected evaluations on the MedMNIST benchmark, we first establish that PEFT outperforms full network updates, improving average accuracy from 86.50% to 88.97%. We then show that a single *MoPET* model consolidates four heterogeneous datasets into one network, improving average accuracy over the best isolated PEFT adapters (93.46% versus 92.83%). Finally, we show that co-training with auxiliary datasets improves accuracy on data-constrained clinical targets, raising average target accuracy over the strongest isolated adapter from 81.58% to 83.58%. Our source code is publicly available at https://github.com/sdoerrich97/mopet.

Keywords: Foundation Models · Parameter-Efficient Fine-Tuning · Mixture of Experts.

1 Introduction

The transition from curated datasets to real-world clinical deployment requires models capable of reasoning across profound anatomical heterogeneity. Generalist foundation models trained on natural images, such as CLIP [25] and DINOv3 [27], provide highly robust visual representations. However, adapting these massive architectures to the nuanced visual domains of specific medical tasks presents a structural dilemma [28]. Conversely, medical-specific foundation models [5,31,37] capture domain-relevant features but typically specialize

* These authors contributed equally to this work.

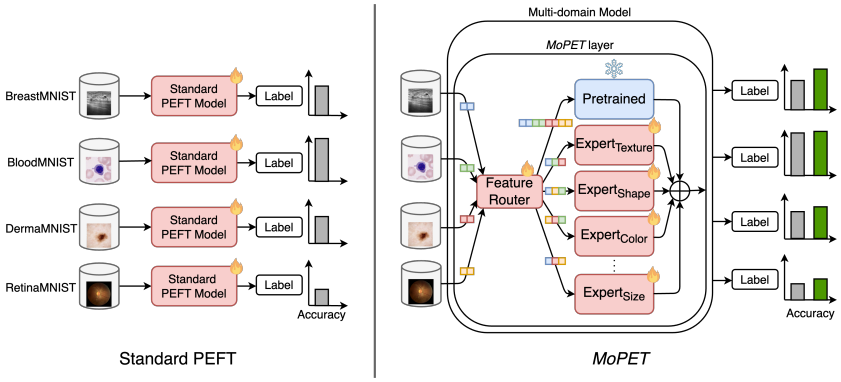

Fig. 1. Left: Standard Parameter-Efficient Fine-Tuning (PEFT) requires independent, disconnected models for distinct medical modalities. Right: *MoPET* dynamically routes diverse anatomical inputs through specialized low-rank experts within a single unified architecture, preventing gradient collision and boosting overall accuracy.

in narrow anatomical regions, demanding necessary adaptation for broader clinical application. Regardless of the base model, full end-to-end finetuning remains computationally prohibitive and consistently overfits in limited data scenarios. Parameter-efficient techniques like Low-Rank Adaptation (LoRA) [12] resolve this computational burden and data scarcity problem by freezing pretrained weights and updating only injected low-rank matrices, requiring substantially fewer training examples. Although LoRA successfully retains generalization, it requires training a separate adapter for every new task or modality. This fragmented strategy spawns a massive proliferation of disparate adapters, which complicates deployment in resource-constrained clinical settings.

To unify these separate adapters into a single network, architectures must overcome the common problem of negative transfer [26]. When optimization gradients from diverse visual domains conflict (such as the high-frequency cellular textures of histopathology versus the low-frequency geometric patterns of ultrasound) model training becomes challenging. Mixture of Experts (MoE) architectures [9,13], including the recent M4oE [15], MedMoE [7], and Med-MoE [14], resolve this interference through sparse routing mechanisms that dynamically assign features to modality-specific experts. Closest to our setting, mixtures of low-rank experts couple sparse routing with parameter-efficient adapters, though primarily in language models and with a homogeneous expert pool [36,18]. Building on these principles, we introduce *MoPET*, a Mixture of Parameter-Efficient Fine-Tuned Experts that condenses distinct medical image classification tasks into a single model. Unlike prior medical MoE frameworks that tie experts to predefined imaging modalities, *MoPET* learns its expert assignments from data, without modality labels at the routing stage. A learned router directs each input through a small subset of a heterogeneous pool of LoRA and BOFT experts, so that domain-specific adaptation and shared semantic reasoning need not com-

pete for the same parameters, while a dynamic minimum-size sampler and per-dataset classification heads address the scale imbalance and disjoint label spaces across datasets. This reduces gradient conflict across domains while retaining the memory efficiency of low-rank adaptation in a single unified model (Fig. 1). Our contributions are as follows:

- We first establish across 12 distinct medical datasets that parameter-efficient finetuning outperforms full end-to-end network updates. This motivates our design choice of building the unified architecture from low-rank adapters rather than full updates.
- We introduce *MoPET*, a novel Mixture of Experts framework built entirely from parameter-efficient modules that condenses distinct medical image classification tasks into a single model. We demonstrate that this outperforms isolated parameter-efficient adapters trained on individual domains for a selected set of four heterogeneous datasets of varying sizes and modalities.
- Finally, we reveal a novel cross-domain training dynamic where the inclusion of auxiliary medical datasets acts as a performance booster. We demonstrate this by elevating the predictive accuracy across three distinct clinical domains through joint training with auxiliary data pools.

2 Methodology

We propose *MoPET*, an architecture that adapts the mixture-of-experts paradigm to parameter-efficient finetuning, for establishing a unified classification model that simultaneously reasons across multiple distinct medical datasets. When trained jointly, diverse medical modalities typically suffer from negative transfer, thereby degrading overall predictive performance. To resolve this, our method integrates 4 sequential components: a dynamic sampling strategy to balance data ingestion, a routing mechanism to isolate conflicting features into specialized low-rank pathways, a load-balancing loss to ensure optimal expert utilization, and task-specific classification heads (Fig. 2).

2.1 Feature Aware Routing Architecture

Standard end-to-end finetuning of large scale networks on medical data frequently leads to overfitting and catastrophic forgetting. To circumvent this, *MoPET* integrates the Mixture of Experts (MoE) paradigm with Parameter Efficient Fine Tuning (PEFT). We maintain a frozen pretrained backbone and inject a small set of trainable low rank matrices (experts) into the query, key, and value projection layers of the self-attention blocks to form a unified adaptation layer. This approach isolates the learning of new domain specific features while preserving the robust generalist representations of the original network. However, simultaneously updating a single set of shared adapters across divergent clinical tasks can cause negative transfer. To physically separate competing gradient updates among the K introduced experts, we deploy a learnable gating network

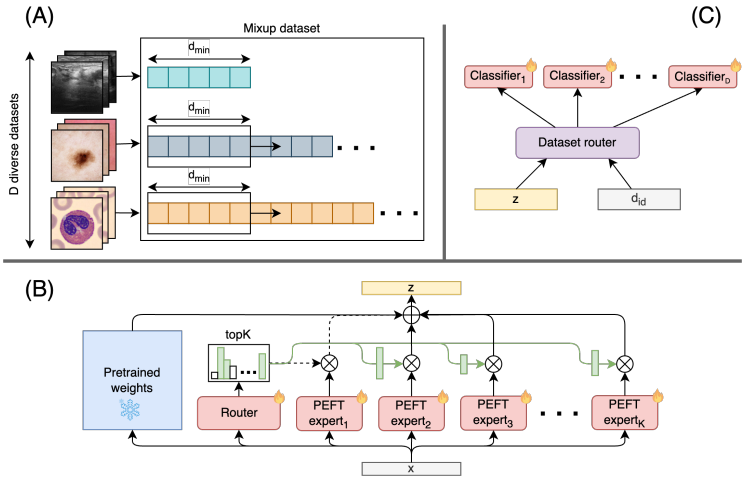

Fig. 2. Overview of *MoPET*. First, a dynamic sampling protocol extracts balanced batches of size d_{min} from each dataset (A). Second, during the forward pass, an internal router activates the top-k parameter-efficient (PEFT) experts to process the input representation $\mathbf{x}$ alongside the frozen pretrained weights, producing the updated token representation $\mathbf{z}$ (B). Finally, a dataset router utilizes the domain identifier d_{id} to direct the terminal representation $\mathbf{z}$ to the corresponding task-specific classifier (C).

R. For a given input token vector $\mathbf{x}$, the gating network $R(\mathbf{x})$ computes raw, unnormalized routing scores to evaluate the relevance of each available expert. We define the final hidden representation $\mathbf{z}$ by augmenting the frozen weight matrix $\mathbf{W}$ with the dynamically weighted expert outputs $\mathbf{e}_i(\mathbf{x})$:

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \sum_{i=1}^{K} \text{softmax}(\text{TopK}(R(\mathbf{x}), k))_i \mathbf{e}_i(\mathbf{x}) \tag{1}$$

The TopK operation retains the k highest routing scores and sets the rest to $-\infty$, so the softmax assigns zero weight to unselected experts and the sum reduces to the k active ones. Each expert $\mathbf{e}_i$ is a parameter-efficient module, either a LoRA or a BOFT adapter, while the frozen projection $\mathbf{W}\mathbf{x}$ acts as an always-on shared path. Routing each token through a subset of experts lets the model adapt to divergent domains while limiting interference between their gradient updates.

To encourage the router to use the full expert pool and avoid expert collapse, where the router repeatedly selects a narrow subset of experts, we add the differentiable load-balancing loss of DeepSeekMoE [9]:

$$\mathcal{L}_{\text{load}} = K \sum_{i=1}^{K} f_i \, p_i, \tag{2}$$

Here, f_i is the fraction of tokens routed to expert i, and p_i is its mean routing probability over the batch. Minimizing this loss promotes a uniform distribution of tokens across all K experts.

We formulate our composite training set as $\mathcal{D} = \{(\mathbf{X}_d, \mathbf{Y}_d)\}_{d=1}^{D}$, where $\mathbf{X}_d$ denotes the input images, $\mathbf{Y}_d$ the corresponding labels, and D indicates the total number of distinct clinical domains. Because this composite dataset $\mathcal{D}$ contains completely divergent label spaces, we attach D independent linear classification heads to the frozen backbone. While the expert layers dynamically route and mix features internally, a dataset router explicitly uses the domain identifier d_{id} to map the shared terminal representations $\mathbf{z}$ strictly into their corresponding semantic spaces, enabling dedicated predictions for every dataset $d \in \{1, \ldots, D\}$. We train the experts, router, and heads jointly by minimizing $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{load}}$, where $\mathcal{L}_{\text{CE}}$ is the cross-entropy on each sample's dataset-specific head.

2.2 Dynamic Multi Domain Sampling

Medical imaging benchmarks frequently exhibit highly skewed class distributions and severe volumetric disparities. If we sample naively, large datasets overwhelm the network and starve smaller domains. To enforce stable optimization across the composite training set $\mathcal{D}$, we deploy a dynamic sliding window sampling protocol. We structure each training epoch to contain exactly $D \times d_{\min}$ samples, where $d_{\min}$ is the number of training samples in the smallest domain. This protocol guarantees the model processes the entirety of the smallest datasets in every single epoch while progressively iterating through the unseen samples of the larger datasets over successive epochs, strictly preventing domain starvation.

2.3 Implementation Details

We instantiate *MoPET* using the DINOv3 [27] Base architecture from the timm library [32] as our frozen backbone, which processes inputs at a spatial resolution of 256×256 pixels. While model performance typically scales with expert count [24], we constrain *MoPET*'s complexity to $K = 32$ experts per Transformer block following established scaling conventions [23], while utilizing the pre-trained weight matrix as a shared expert [9]. For the specialized pathways, we use 20 Low-Rank Adaptation (LoRA) experts with a rank $r = 8$ and a scaling factor $\alpha = 8$, alongside 12 BOFT (Butterfly Orthogonal Fine-Tuning) experts utilizing a block size of 8 and a butterfly factor of 1. During isolated parameter-efficient finetuning, LoRA achieves the highest overall accuracy, while BOFT surpasses it on a subset of datasets (notably Breast and Retina). This complementarity motivates our hybrid expert pool, with more experts allocated to the stronger LoRA family. A sparse gating mechanism activates the top-$k = 12$ experts per forward pass. In total, *MoPET* trains 7.4M parameters (≈8.7% of the 86M backbone). While a single isolated adapter is lighter (LoRA 0.30M, BOFT 0.10M), it serves only one task; *MoPET* instead unifies all tasks in one model, replacing a per-task collection of adapters and the model zoo it entails. Finally, we attach task-specific classification heads for each distinct dataset. We conduct training for 75 epochs with a batch size of 128 using AdamW [22] (initial learning rate of 1×10^{-3}), cosine annealing scheduling [21], early stopping based on validation performance (patience of 10 epochs) and a loss weight $\lambda = 0.05$.

3 Experiments and Results

We evaluate on twelve 2D datasets from the MedMNIST+ collection [34,35] (2–11 classes; CC BY 4.0 / CC BY-NC 4.0), provided at 224×224 pixels and resized to 256×256 with bicubic interpolation to match the backbone. We follow the official data splits throughout. They cover eight imaging modalities across a wide range of scales: blood cell microscopy (Blood [1]), breast ultrasound (Breast [2]), chest X-ray (Chest [30], Pneumonia [17]), dermatoscopy (Derma [29,8]), retinal OCT (OCT [17]), abdominal CT in axial, coronal, and sagittal views of the same volumes (OrganA, OrganC, OrganS [3,33], treated as distinct tasks), colon pathology (Path [16]), kidney cortex microscopy (Tissue [34]), and fundus photography (Retina [19]). Training sets span from 546 images (Breast) to roughly 165,000 (Tissue), a scale heterogeneity that motivates our dynamic multi-domain sampler.

3.1 Isolating Adaptation Strategy from Backbone Influence

First, we establish that parameter-efficient fine-tuning outperforms full end-to-end network updates. To isolate the influence of the backbone from the fine-tuning method, we compare CLIP [25], DINO [4], and DINOv3 [27] under full end-to-end fine-tuning and the PEFT techniques LoRA [12], BOFT [20], FourierFT [11], and AdaptFormer [6]. Table 1 reports the mean accuracy, evaluated at an operating point of 0.5, across the test splits of all twelve datasets and three seed runs. The end-to-end results for DINO and CLIP, as well as the DINOv3 training procedure, are adapted from [10]. For PEFT, we adjusted the learning rate from 0.0001 to 0.001. When evaluating the backbones under parameter efficient settings, the architectures exhibit high performance stability. A Friedman test reveals no statistically significant differences among the foundation models ($\chi^2 = 3.63$, $p = 0.163$, $\alpha = 0.05$). While a non-significant test does not establish equivalence, backbone selection alone does not provide reliable performance separation in clinical tasks.

Conversely, the method of parameter update dictates diagnostic success. On aggregate across the evaluated datasets, PEFT methods, specifically LoRA and BOFT, outperform traditional full network updates. Pairwise Wilcoxon signed-rank tests with Bonferroni correction confirm this: both LoRA and BOFT significantly outperform end-to-end fine-tuning ($p \leq 1.2 \times 10^{-3}$) and AdaptFormer ($p < 10^{-8}$). LoRA additionally outperforms FourierFT ($p = 4.7 \times 10^{-7}$), and BOFT likewise outperforms FourierFT ($p = 3.7 \times 10^{-5}$), whereas LoRA and BOFT are statistically indistinguishable ($p = 0.08$) and end-to-end fine-tuning and FourierFT do not differ significantly ($p = 1.00$).

3.2 Cross-Dataset Adaptation

Next, we evaluate the cross-dataset adaptation capabilities of *MoPET*. For this, we train a single, unified model simultaneously on a selected pool of four heterogeneous datasets of varying sizes and modalities (BloodMNIST, BreastMNIST,

Table 1. Average test accuracy in %, evaluated at an operating point of 0.5, of different PEFT methods and backbones across all twelve MedMNIST datasets and three seed runs in comparison to end-to-end fine-tuning. Best backbone per fine-tuning method is shown in bold (columns), while best fine-tuning method per backbone is underlined (rows).

Model	LoRA	BOFT	FourierFT	AdaptFormer	endToEnd
CLIP	88.04±0.87	87.37±0.92	84.57±0.61	**82.84±1.74**	82.75±1.01
DINO	88.14±0.81	87.28±1.19	85.43±0.60	82.08±0.96	84.84±1.07
DINOv3	**88.97±0.73**	**88.62±0.68**	**86.72±0.21**	79.82±1.13	**86.50±1.15**

DermaMNIST, and PathMNIST) and compare it against the individually trained baseline adapters using DINOv3 as backbone. In this multi-domain configuration, our dynamic sampling protocol restricts the number of samples per dataset per epoch to $d_{\min} = 546$, matching the training size of the smallest domain (BreastMNIST). This mechanism prevents the larger datasets from dominating the gradient updates, thereby ensuring balanced anatomical representation. Table 2 reports test accuracy across the four test splits for the individually trained baselines and our single *MoPET* model, as the mean over three seeds. Compared to the strongest isolated PEFT adapter, *MoPET* improves on BreastMNIST, DermaMNIST, and PathMNIST and stays within 0.11% on BloodMNIST, raising the average from 92.83% (BOFT) to 93.46%. End-to-end fine-tuning remains strongest on PathMNIST but trails on the smaller, more heterogeneous targets. That a single model matches or exceeds the best per-dataset adapter on three of four datasets indicates that the routed experts share features across domains rather than overfitting to dataset-specific cues.

Table 2. Test accuracy in %, evaluated at an operating point of 0.5, comparing traditional fine-tuning and isolated parameter-efficient adapters against the unified *MoPET* framework using the DINOv3 backbone, on a selected pool of four heterogeneous datasets of varying sizes and modalities (BloodMNIST, BreastMNIST, DermaMNIST, and PathMNIST). All values are the mean over three random seeds. The best method per dataset is underlined; *MoPET* is shown in bold where it ranks among the top three methods.

Method	BloodMNIST	BreastMNIST	DermaMNIST	PathMNIST	Average
End-to-End	98.49	85.26	81.23	96.13	90.28
Linear Probing	97.86	86.54	82.28	93.90	90.15
AdaptFormer	96.50	73.29	69.89	89.06	82.19
BOFT	98.81	90.38	86.93	95.21	92.83
FourierFT	98.40	88.46	83.18	94.52	91.14
LoRA	98.75	88.25	87.43	95.07	92.38
MoPET	**98.70**	**90.81**	**88.46**	**95.87**	**93.46**

Table 3. Test accuracy in % between traditional fine-tuning, isolated parameter-efficient adapters, and the *MoPET* boosting setup using the DINOv3 backbone. *MoPET* boosts each primary target (BreastMNIST, RetinaMNIST, DermaMNIST) with a different, hand-selected pool of auxiliary datasets during training. All values are the mean over three random seeds; the best method per target is underlined.

	BreastMNIST	RetinaMNIST	DermaMNIST	Average
End-to-End	85.26	55.25	81.23	73.91
Linear Probing	86.54	65.17	82.28	78.00
LoRA	88.25	66.83	87.43	80.84
BOFT	90.38	67.42	86.93	81.58
MoPET	<u>92.95</u>	<u>68.83</u>	<u>88.96</u>	<u>83.58</u>

3.3 Cross-Dataset Feature Sharing via Auxiliary Booster Datasets

Finally, we investigate whether auxiliary datasets can further boost the predictive accuracy of individual target domains during *MoPET*'s joint training. To evaluate this paradigm, we select three distinct booster configurations. Rather than adapting the DINOv3 backbone exclusively to a single dataset, we jointly train it alongside additional support datasets. Specifically, we co-train BreastMNIST with Blood-, Derma-, and PathMNIST; RetinaMNIST with Blood-, Breast-, Path-, and OrganAMNIST; and DermaMNIST with Blood-, OCT-, and OrganSMNIST. Table 3 reports test accuracy on the three target datasets as the mean over three seeds. For reference, we include traditional fine-tuning and the isolated adapters. With its auxiliary pools, *MoPET* reaches 92.95% on Breast-, 68.83% on Retina-, and 88.96% on DermaMNIST, for an average of 83.58%, ahead of every isolated baseline. The gains are largest on the smallest targets, Breast- and RetinaMNIST (546 and 1,080 training images), indicating that co-training with additional data benefits data-constrained domains the most.

4 Discussion and Conclusion

We introduce *MoPET*, a mixture-of-experts method that couples a learned sparse router with low-rank experts so that a single frozen foundation model can serve many medical classification tasks at once. Across twelve datasets we first establish that parameter-efficient fine-tuning outperforms full end-to-end network updates, which motivates building the unified model from low-rank adapters. We then show that one *MoPET* model consolidates four heterogeneous datasets and, on average, exceeds the best per-dataset adapter. Finally, co-training with auxiliary datasets improves accuracy on data-constrained targets, with the largest gains on the smallest datasets. The same joint formulation thus serves two ends, a single unified model or a boosted individual target.

Limitations. Several questions remain open. Key design choices, notably the expert count K, the top-k activation, and the LoRA-to-BOFT ratio, are likely dataset dependent and warrant a more rigorous ablation than our scope allows. We also do not directly analyze the learned routing patterns; whether experts specialize by modality or anatomy, as intended, remains to be verified through an explicit study of expert utilization. Our cross-dataset and booster experiments further rely on hand-selected dataset subsets chosen to span modalities and scales; a systematic sweep over combinations, together with control conditions using randomly selected or unrelated auxiliary data, would be needed to separate semantically driven transfer from the effect of added training data. Finally, we do not benchmark against dedicated multi-task medical pretraining approaches, which would form an interesting comparison for the cross-dataset feature-sharing setting.

Acknowledgments. This study was funded through the Hightech Agenda Bayern (HTA) of the Free State of Bavaria, Germany.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acevedo, A., Merino, A., Alférez, S., Ángel Molina, Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in Brief **30**, 105474 (2020)
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020)
3. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
4. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9630–9640 (2021)
5. Chen, R.J., Ding, T., Lu, M.Y., et al.: Towards a General-Purpose Foundation Model for Computational Pathology. Nature medicine **30**(3), 850–862 (Mar 2024)
6. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems (NeurIPS) **35** (2022)
7. Chopra, S., Sanchez-Rodriguez, G., Mao, L., Feola, A.J., Li, J., Kira, Z.: MedMoE: Modality-Specialized Mixture of Experts for Medical Vision-Language Understanding (Jun 2025)
8. Codella, N.C.F., Gutman, D., Celebi, M.E., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 168–172 (2018)
9. Dai, D., Deng, C., Zhao, C., et al.: DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models (Jan 2024)

10. Doerrich, S., Di Salvo, F., Brockmann, J., Ledig, C.: Rethinking model prototyping through the MedMNIST+ dataset collection. Scientific Reports **15**(1), 7669 (Mar 2025)
11. Gao, Z., Wang, Q., Chen, A., Liu, Z., Wu, B., Chen, L., Li, J.: Parameter-efficient fine-tuning with discrete fourier transform. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (Oct 2021)
13. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive Mixtures of Local Experts. Neural Computation **3**(1), 79–87 (Feb 1991)
14. Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z.: Med-MoE: Mixture of Domain-Specific Experts for Lightweight Medical Vision-Language Models (Sep 2024)
15. Jiang, Y., Shen, Y.: M4oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15012. Springer Nature Switzerland (2024)
16. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLOS Medicine **16**, e1002730 (2019)
17. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**, 1122–1131.e9 (2018)
18. Komatsuzaki, A., Puigcerver, J., Lee-Thorp, J., Ruiz, C.R., Mustafa, B., Ainslie, J., Tay, Y., Dehghani, M., Houlsby, N.: Sparse upcycling: Training mixture-of-experts from dense checkpoints. In: International Conference on Learning Representations (ICLR) (2023)
19. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns **3**, 100512 (2022)
20. Liu, W., Qiu, Z., Feng, Y., Xiu, Y., Xue, Y., et al.: Parameter-efficient orthogonal finetuning via butterfly factorization. In: ICLR (2024)
21. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts (2017)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019)
23. OpenAI, Agarwal, S., Ahmad, L., et al.: Gpt-oss-120b & gpt-oss-20b Model Card (Aug 2025)
24. Pan, D., Li, B., Zheng, Y., Ma, J., Fei, V.: The Rise of Sparse Mixture-of-Experts: A Survey from Algorithmic Foundations to Decentralized Architectures and Vertical Domain Applications (Feb 2026)
25. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
26. Ruder, S.: An Overview of Multi-Task Learning in Deep Neural Networks (Jun 2017)
27. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., et al.: DINOv3 (Aug 2025)
28. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (Sep 2022)

29. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 2018 5:1 **5**, 1–9 (2018)
30. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3462–3471 (2017)
31. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text (Oct 2022)
32. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019)
33. Xu, X., Zhou, F., Liu, B., Fu, D., Bai, X.: Efficient multiple organ localization in ct image using 3d region proposal network. IEEE Transactions on Medical Imaging **38**(8), 1885–1898 (2019)
34. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data **10**(1), 41 (Jan 2023)
35. Yang, J., Shi, R., Wei, D., Liu, Z., et al.: MedMNIST+: 18× standardized datasets for 2d and 3d biomedical image classification with multiple size options: 28 (mnist-like), 64, 128, and 224 (2024)
36. Zadouri, T., Üstün, A., Ahmadian, A., Ermiş, B., Locatelli, A., Hooker, S.: Pushing mixture of experts to the limit: Extremely parameter efficient MoE for instruction tuning. In: International Conference on Learning Representations (ICLR) (2024)
37. Zhang, S., Xu, Y., Usuyama, N., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (Jan 2025)