Evaluating Self-Supervised Learning in Medical Imaging: A Systematic Investigation of Robustness, Generalizability, and Multi-Domain Impact

Valay Bundele*
valaybundele@gmail.com
Karahan Sarıtaş*
karahansaritas@gmail.com
Bora Kargi*
kargibora@gmail.com
Oğuz Ata Çal*
ouzal@rocketmail.com
Kıvanç Tezören*
kivanctezoren@gmail.com
Zohreh Ghaderi
zohreh.ghaderi@uni-tuebingen.de
Hendrik Lensch
hendrik.lensch@uni-tuebingen.de

Institute for Computer Graphics
University of Tuebingen
Tuebingen, Germany

Abstract

Self-supervised learning (SSL) has emerged as a promising paradigm in medical imaging, addressing the chronic challenge of limited labeled data in healthcare settings. While SSL has shown impressive results, existing studies in the medical domain are often limited in scope, focusing on specific datasets or modalities, or evaluating only isolated aspects of model performance. This fragmented evaluation approach poses a significant challenge, as models deployed in critical medical settings must not only achieve high accuracy but also demonstrate robust performance and generalizability across diverse datasets and varying conditions. To bridge this gap, we conduct a rigorous investigation into the design space of SSL for medical imaging, evaluating 8 major SSL methods across 11 real-world medical datasets. Our analysis spans three core dimensions: (1) in-domain performance under varying label proportions (1%, 10%, and 100%), (2) cross-dataset generalization, and (3) robustness to out-of-distribution (OOD) samples. Beyond empirical evaluation, we further examine how initialization strategies, model architectures, and multi-domain pre-training contribute to SSL's success in medical imaging.

*These authors contributed equally to the work.

1 Introduction

Medical image annotation is costly and labor-intensive due to the need for expert knowledge [44, 50, 69, 85]. This annotation bottleneck limits the scalability of supervised learning in healthcare, motivating the rise of self-supervised learning (SSL) as a powerful alternative. SSL leverages large-scale unlabeled data to learn generalizable representations, enabling downstream tasks with minimal labeled supervision [2, 7, 24, 27, 43, 45, 77].

While SSL has shown promise in improving classification with limited labels [15, 16, 17, 32, 86], its robustness and generalizability remain underexplored, particularly in the medical domain. Clinical deployment demands models that not only generalize across diverse datasets and modalities but also recognize out-of-distribution (OOD) samples where predictions should be withheld. Existing OOD studies in SSL mostly focus on natural images [34, 40, 48, 53, 62], with limited work in the medical domain. Prior studies typically target a single modality with supervised techniques [10], evaluate few SSL methods in one domain [12], or focus on task-specific rather than general-purpose SSL [13].

To address these gaps, we conduct a systematic investigation into the SSL design space in medical imaging, evaluating robustness and generalizability across diverse settings. We assess various SSL methods for OOD detection, comparing the effectiveness of convolutional networks with transformer-based architectures. Moreover, we perform cross-dataset transfer evaluations to rigorously assess the generalizability of learned representations.

Transfer learning—initializing models with IMAGENET1K pre-trained weights—is widely used in medical imaging [23, 61, 63, 74]. These weights are either directly fine-tuned or used to initialize SSL pre-training [24, 60]. We adopt the latter, enabling the model to adapt better to the fine-grained medical data. Moreover, we analyze how IMAGENET1K initialization affects generalization and OOD performance, an underexplored aspect in prior work.

Recent work suggests that training on multiple domains improves model robustness, particularly in data-limited and OOD-prone medical scenarios [66, 81, 82]. Multi-domain training can leverage complementary information across datasets [18, 76], but the role of multi-domain SSL remains underexplored. To address this, we investigate how multi-domain pre-training influences generalization and robustness across methods and architectures.

To summarize, we explore the following key questions:

Q1. How does the SSL method, initialization strategy, and architecture affect in-domain performance, generalizability, and robustness to OOD?

Q2. Which SSL methods and initializations perform best under limited labeled data?

Q3. Does multi-domain training enhance the in-domain performance, robustness and generalizability of SSL techniques?

By addressing these questions, we aim to provide a clearer understanding of SSL's potential in building reliable, generalizable models for real-world medical applications.

2 Related Work

SSL in Medical Imaging While SSL has been extensively benchmarked on natural images [21, 30, 59], evaluations in medical domain remain limited. Doerrich et al. [24] focused on IMAGENET1K pre-trained encoders without domain-specific fine-tuning. Kang et al. [45] showed SSL's advantage over supervised baselines in pathology, while Afzal et al. [3] evaluated several SSL methods for multi-view chest X-ray classification, but focused on a single modality. Huang et al. [43] compared SSL and semi-supervised methods on four medical datasets, without assessing robustness or generalizability. In contrast, we provide a comprehensive evaluation across several medical datasets, focusing on multi-domain performance, robustness, generalizability, and adaptability to limited labeled data.

Out-of-Distribution Detection SSL has shown promise for OOD detection in natural images [40, 53, 62]. In medical domain, prior work mainly relies on supervised training [64], investigates only task-specific SSL methods [14] or evaluates only a few SSL methods on a single modality [70]. To our knowledge, no prior work has systematically examined several SSL methods for OOD detection in medical domain, across diverse architectures, initializations and pretraining strategies, a gap we address.

Generalizability Generalization across domains is critical in medical imaging. Previous studies [28, 52, 79] offer partial views, often constrained to specific modalities or a few SSL methods. Azizi et al. [8] assess generalizability to OOD shifts from devices or clinical demographics. DiRA [36] introduces a collaborative learning strategy and evaluates generalization but focuses solely on the proposed method. In contrast, we present a systematic cross-dataset evaluation, analyzing the effects of initializations, architectures, and multi-domain training to provide a comprehensive assessment of generalization.

Transfer Learning Transfer from IMAGENET1K remains common in medical imaging [8, 60, 73, 74]. Although the in-domain benefits of continual pre-training where IMAGENET1K weights are used to initialize SSL pretraining are known [41, 61], its OOD detection and generalizability impact is underexplored. In this work, we investigate whether continued pre-training improves both generalizability and OOD detection performance. Moreover, we are the first to explore self-supervised IMAGENET1K initialization for SSL in medical imaging.

3 Methodology

Representation Learning Methods We consider the following eight discriminative SSL methods: SimCLR [16], DINO [15], BYOL [31], ReSSL [86], MoCo v3 [17], NNCLR [26], VICREG [9], and Barlow Twins [83] which are explained briefly in Appendix B. A comprehensive survey by Huang et al. [42] highlights SimCLR, MoCo, and BYOL as the most frequently adopted SSL frameworks in medical imaging research. We include the other methods to make our study more comprehensive.

Tasks and Datasets The datasets utilized in this study are derived from MedMNIST [80], which was introduced to standardize research on medical imaging tasks. Its design enables reproducible, controlled and rapid evaluations. It has been widely adopted in medical SSL [43, 65]. To maintain our focus on multiclass classification and ordinal regression, we exclude ChestMNIST from our analysis, as it only offers multi-label classification. Our experiments span 11 MedMNIST+ datasets, additional information is provided in Appendix A.

Architectures For our study, we use ResNet-50 [37] with approximately 25 million parameters and ViT-Small [25] with approximately 22 million parameters to ensure a fair comparison. We exclude other models like ViT-Tiny and ViT-Base due to their significantly different parameter counts of 5.7 million and 86 million, respectively.

Pre-training We employed `solo-learn` library [21] with some modifications for pre-training, OOD detection and linear evaluation. Our implementation details are provided in Appendix C.1, and code will be released after acceptance. In total, we focus on five different pre-training schemes: supervised training with (1) random initialization and (2) supervised IMAGENET1K initialization; self-supervised pre-training with (3) random initialization, (4) supervised IMAGENET1K, and (5) self-supervised IMAGENET1K initializations.

Linear Evaluation We assess the quality of SSL encoders via linear probing, training logistic regression classifiers on frozen features [31, 38, 49]. To reflect realistic medical settings, we evaluate low-shot performance using 1% and 10% labeled data, following Caron et al. [15]. We report accuracy and AUC scores. Full details are in Appendix C.3 and C.4.

Generalizability To assess generalizability, we conduct cross-dataset experiments using our dataset collection, $\mathcal{D} = \{D_1, ..., D_{11}\}$. We pre-train each SSL method on a dataset D_i and evaluate its transferability by training a linear classifier on frozen features on all other datasets, $\mathcal{D} \setminus D_i$. This process is repeated for each D_i in $\mathcal{D}$ to cover all cross-dataset pairs.

OOD Detection We evaluate various SSL methods for distinguishing between ID and OOD samples, focusing on backbone architectures, pre-training strategies, and multi-domain learning. We follow the approach defined in [51], where feature representations $\mathbf{f}(x) \in R^d$ are assumed to follow a class-conditional multivariate Gaussian, with parameters estimated from the ID training data $\mathcal{P}_{\text{ID}}$ (i.e., samples used to train the encoder). For each class c, the Mahalanobis distance is computed as $D_{\text{M}}(x,c) = \sqrt{(\mathbf{f}(x) - \mu_c)^\top \Sigma^{-1} (\mathbf{f}(x) - \mu_c)}$ and the confidence score is given by $S(x) = -\min_c D_{\text{M}}(x,c)$.

Samples are considered OOD if they originate from a different dataset or subset (denoted as $\mathcal{P}_{\text{OOD}}$) than the one used for training. OOD detection performance is evaluated using threshold-independent metrics such as AUROC and AUPR. Further details are in Appendix.

Multi-Domain Learning We examine the impact of multi-domain learning by comparing models trained on single-domain MedMNIST datasets with those trained on two combinations. The first combination, named Organ{A,C,S} merges the OrganAMNIST, OrganCMNIST and OrganSMNIST datasets to represent a single-modality scenario. The second combination, referred to as Organ{A,S}PnePath, includes OrganAMNIST (CT), OrganSMNIST (CT), PathMNIST (Colon Pathology), and PneumoniaMNIST (Chest X-Ray), allowing us to assess the effect of combining different modalities. We then train SSL methods from scratch on these combined datasets to assess how dataset combination affects performance.

4 Experiments & Results

In this section, we analyze the in-domain (ID) performance, robustness to OOD samples, and generalizability of the SSL methods. For clarity in certain figures where direct comparisons between the SSL methods are not essential, we display results for a representative subset of five methods, with complete results available in the Appendix for reference.

4.1 In-Domain Performance

We evaluate in-domain performance by pre-training ResNet-50 with each SSL method and training a linear classifier on frozen features. MoCo v3 achieves the highest accuracy on 5 of 11 datasets, followed closely by BYOL and SimCLR. Notably, SSL outperforms supervised learning on 7 of 11 datasets when both use random initialization (Appendix D.1).

We then analyze the impact of supervised IMAGENET1K initialization. As shown in Figure 1, IMAGENET1K initialization consistently boosts in-domain performance, with DINO and BYOL showing the largest gains. This trend is consistent across methods (Appendix D.1).

Figure 2 highlights the accuracy gap between ResNet-50 and ViT ($\text{ACC}_{\text{RN50}} - \text{ACC}_{\text{ViT}}$) under both random and IMAGENET1K initialization. ResNet-50 generally outperforms ViT, but the gap narrows with IMAGENET1K initialization. This is likely due to diminishing gains at higher accuracy and the improved data efficiency of transformers when initialized with large-scale pre-training [4, 6, 25].

Expanding on our in-domain performance evaluation, we also examine how the architectures perform under different levels of label availability. Table 1 presents accuracy drops for various SSL methods when reducing label availability from 100% to 1% in random initialization setting. ResNet-50 not only shows better performance in the fully labeled (100%) scenario, as shown in Figure 2, but also exhibits consistently lower average performance drops across nearly all methods. This indicates that ResNet-50 maintains more robust performance compared to ViT-Small even under severe label scarcity. A similar pattern can also be observed across other methods and in IMAGENET1K initialization setting in Appendix D.1, highlighting that transformers generally require more labeled data [77, 87].

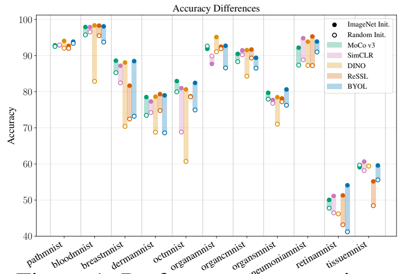


Figure 1: Performance comparison of random (unfilled) vs. imagenet (filled) init. using ResNet-50.

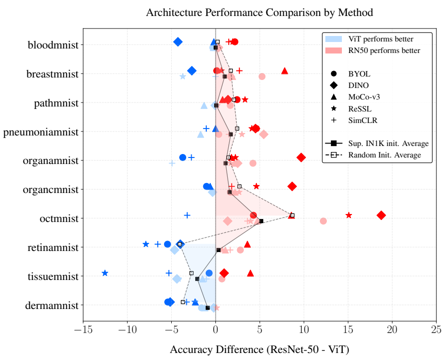


Figure 2: Performance difference between ResNet-50 and ViT across SSL methods and initialization strategies. Dark and light colors indicate random and ImageNet initialization, respectively.

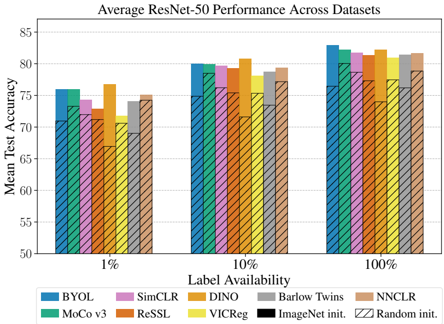
Moreover, Figure 3(a) presents a bar chart comparing mean test accuracies of ResNet-50 across different label availability settings (1%, 10%, and 100%), highlighting the effects of IMAGENET1K versus random initialization. The results show that models initialized with IMAGENET1K weights consistently outperform their randomly initialized counterparts, especially under label-scarce conditions. This finding underscores the value of pre-trained weights in boosting model performance when labeled data is limited. Notably, with IMAGENET1K initialization, DINO excels in label-scarce scenarios, though it is slightly outperformed by BYOL at the 100% label level. On the other hand, MoCo v3 shows promising results across most of the label percentages in the random initialization setting, slightly losing only to NNCLR at the 1% label level. A similar trend for ViTs is shown in Appendix D.1.

Next, we examine the impact of using self-supervised IMAGENET1K initialization, as opposed to supervised IMAGENET1K initialization, on in-domain performance using MoCo v3, SimCLR, BYOL, and DINO. A paired t-test comparing the two initializations across all datasets and methods, using normalized accuracies to account for varying dataset difficulty levels, revealed no statistically significant difference in performance ($t(43) = 1.167$, $p = 0.2496$, mean normalized difference = 0.26σ, 95% CI: [-0.19σ, 0.70σ]). Individual accuracy scores are in Appendix D.1.

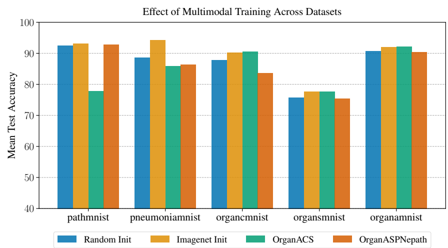
Finally, we examine the impact of multi-domain training on in-domain performance. As shown in Figure 3(b), pre-training on a multi-domain dataset consisting of similar domains, (Organ{A,C,S}) with random initialization (green bars) outperforms training on constituent single-domain datasets with random initialization (blue bars). Furthermore, it achieves per-

Table 1: Accuracy drops from 100% to 1% label availability across datasets (random init). Larger drops are highlighted in red, emphasizing higher sensitivity to limited labeled data.

	SimCLR		DINO		BYOL		ReSSL		MoCo v3	
	RN50	ViT	RN50	ViT	RN50	ViT	RN50	ViT	RN50	ViT
Path	2.91	0.10	0.24	1.22	1.10	1.60	2.38	0.22	-0.12	1.50
Derma	4.54	7.93	1.90	6.56	1.68	4.51	5.54	8.41	3.00	5.47
OCT	9.24	5.62	0.56	6.08	-2.24	7.52	-2.16	10.76	4.20	8.18
Pneumonia	-0.13	9.84	-2.85	0.22	0.06	-2.21	-3.07	3.72	0.45	-0.67
Retina	4.50	9.65	3.30	5.85	-0.85	1.90	0.40	7.00	4.70	0.60
Breast	7.82	3.72	0.51	0.00	0.13	0.00	0.77	1.15	6.54	0.90
Blood	3.34	13.11	12.34	14.57	8.45	17.43	9.92	13.91	6.28	12.04
Tissue	6.66	6.37	7.45	5.04	7.10	8.69	5.76	6.58	7.18	10.20
OrganA	9.40	9.10	8.72	11.73	9.42	6.99	7.36	8.00	6.17	4.63
OrganC	11.45	16.95	23.87	16.93	18.81	15.09	20.32	16.23	14.36	11.20
OrganS	13.83	15.71	21.52	20.16	16.77	15.38	20.13	22.46	21.66	18.33
Average	6.69	8.92	7.05	8.03	5.49	6.99	6.12	8.95	6.76	6.58
Counts	5	6	4	7	5	6	2	10	6	5



(a)



(b)

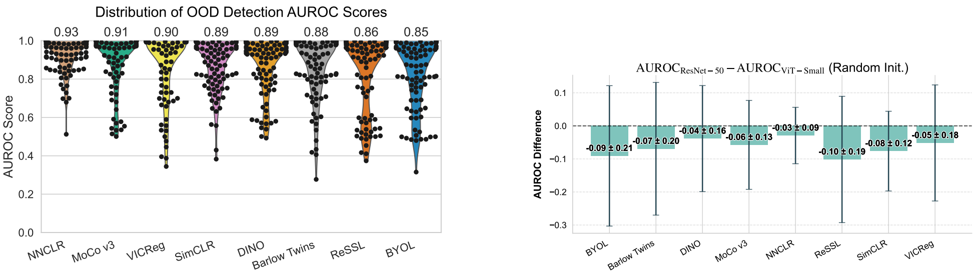
Figure 3: (a) Mean test accuracy across datasets at 1%, 10%, and 100% label availability, comparing IMAGENET1K and random (hatched) initialization using ResNet-50 (b) Mean ID performance of SSL methods (ResNet-50), trained on single-domain datasets with different initialization strategies and multi-domain datasets with random initialization.

formance comparable to single-domain training with IMAGENET1K-initialization (light orange). In contrast, combining datasets from diverse domains in Organ{A,S}PnePath does not significantly improve accuracy, and may even result in poorer performance. These findings suggest that multi-domain pretraining is beneficial when domains of combined datasets are similar whereas combining diverse domains may not benefit in-domain performance.

4.2 Out-of-Distribution Detection

We evaluate OOD detection performance by designating one of the datasets as in-distribution (ID) and the others as out-of-distribution (OOD), yielding 110 tests (11×10) per method. Each SSL encoder is pre-trained on the ID dataset and assessed for its ability to identify OOD samples. Figure 4(a) shows AUROC distributions for OOD detection using randomly initialized ResNet-50 backbones. NNCLR and MoCo v3 achieve the highest scores, suggesting that they are particularly effective in learning representations that differentiate ID from OOD samples. With a ViT-Small backbone, MoCo v3 showed strong OOD detection performance, which aligns with expectations as it was originally developed for ViT architecture. Detailed analysis of effect of backbone choice on each method and in-distribution, OOD pairs $(\mathcal{P}_{\text{ID}}, \mathcal{P}_{\text{OOD}})$ can be found in Appendix D.2.2.

In our experimental setup, the choice of backbone architecture played a crucial role in OOD detection performance. Specifically, models using ViT-Small consistently outperformed those using ResNet-50, as shown in Figure 4(b). This observation aligns with prior research, such as [29, 84], which demonstrated that Transformer-based architectures like ViT generally excel in OOD detection tasks. Furthermore, this trend held across all evaluated SSL methods, suggesting that the improved AUROC scores result from the ViT backbone itself rather than any specific SSL method.

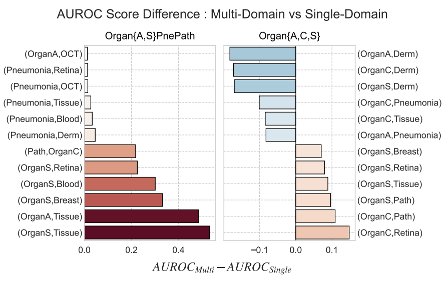


(a) (b)

Figure 4: (a) AUROC distributions for ResNet-50 models (random init). Black dots show individual $(\mathcal{P}_{\text{ID}}, \mathcal{P}_{\text{OOD}})$ AUROC scores, and the mean AUROC score for each method is shown above. (b) Mean AUROC differences between ResNet-50 and ViT-Small across all $(\mathcal{P}_{\text{ID}}, \mathcal{P}_{\text{OOD}})$ pairs. Negative values favor ViT-Small. All differences are statistically significant ($p < 0.05$), demonstrating a clear overall advantage of ViT-Small models.

We observe that the impact of initialization on OOD detection varies depending on backbone, SSL method, and in-domain data $\mathcal{P}_{\text{ID}}$. Generally, for smaller datasets such as BreastMNIST, RetinaMNIST, and DermaMNIST, models initialized with IMAGENET1K weights demonstrate improved performance. For the ResNet-50 backbone, SSL methods such as SimCLR, ReSSL, MoCo v3, and VICReg enhanced OOD detection performance with random initialization. In the case of ViT-Small backbone models, although differences in OOD detection scores were less pronounced, there was still a noticeable tendency for models to favor random initialization, with VICReg showing the strongest preference.

Moreover, initializing models with self-supervised IMAGENET1K weights led to significant differences in AUROC scores compared to starting with supervised IMAGENET1K weights for several datasets. In particular, significant negative differences (i.e., lower AUROC with self-supervised initialization) were observed for BreastMNIST (ΔAUC $= -0.12$, 95% CI: ± 0.18), PathMNIST (ΔAUC $= -0.09$, 95% CI: ± 0.13), and RetinaMNIST (ΔAUC $= -0.03$, 95% CI: ± 0.09). Conversely, significant positive differences (i.e., higher AUROC with self-supervised initialization) were found for PneumoniaMNIST (ΔAUC $= 0.08$, 95% CI: ± 0.15) and TissueMNIST (ΔAUC $= 0.08$, 95% CI: ± 0.19). Detailed information about initialization preferences for each SSL method and backbone are in Appendix D.2.3.



Lastly, we compare OOD detection performance between models trained on multi-domain datasets versus their constituent single-domain datasets. Figure 5 shows that for all SSL methods, Organ{A,S}PnePath consistently improves AUROC scores across most $(\mathcal{P}_{\text{ID}}, \mathcal{P}_{\text{OOD}})$ pairs, with the exception of a single pair. In contrast, Organ{A,C,S} generally decreases OOD detection performance across most $(\mathcal{P}_{\text{ID}}, \mathcal{P}_{\text{OOD}})$ pairs, particularly for models trained with MoCo v3, SimCLR, and ReSSL. These findings demonstrate that a multi-domain dataset with greater domain diversity (Organ{A,S}PnePath) provides a more substantial OOD detection performance boost compared to a more homogeneous multi-domain dataset (Organ{A,C,S}).

4.3 Generalizability

In this section, we assess transferability of SSL representations via cross-dataset evaluations. Specifically, we first pre-train the encoder on one dataset and then train a linear classifier on each of the remaining datasets individually, using a frozen encoder. Generalizability is measured by relative accuracy drop: $\Delta\text{ACC} = \frac{\text{ACC}_{\text{ID}} - \text{ACC}_{\text{transfer}}}{\text{ACC}_{\text{ID}}} \times 100$, where ACC_{ID} represents accuracy when training and testing are performed on the same dataset, and $\text{ACC}_{\text{transfer}}$ denotes accuracy when model is pre-trained on a different dataset. For each source dataset, we report average drop ($\overline{\Delta\text{ACC}}$) across all target datasets, revealing how well learned representations generalize and which datasets yield the most transferable features. Table 2 shows that SimCLR and MoCo v3 excel in generalization, with both achieving the lowest accuracy drops in 4 datasets. Notably, models pre-trained on Path or Organ{C,S}MNIST show the lowest accuracy drops, suggesting these datasets offer the most transferable representations.

Table 2: Average accuracy drop (%) relative to in-domain ResNet-50 (random init), across all targets. Lowest accuracy drops shown in green.

	Accuracy Drop Δ_{ACC} (%)				
Source	SimCLR	MoCo v3	Barlow	NNCLR	ReSSL
Path	$6.44_{\pm 5.78}$	$6.38_{\pm 5.99}$	$3.32_{\pm 5.82}$	$6.80_{\pm 6.41}$	$10.10_{\pm 7.54}$
Derma	$11.16_{\pm 7.00}$	$16.06_{\pm 8.55}$	$13.32_{\pm 8.77}$	$19.81_{\pm 10.60}$	$14.08_{\pm 10.36}$
OCT	$14.18_{\pm 5.06}$	$8.62_{\pm 4.87}$	$8.05_{\pm 4.52}$	$11.40_{\pm 5.68}$	$12.87_{\pm 6.67}$
Pneumonia	$11.46_{\pm 7.84}$	$9.77_{\pm 8.73}$	$17.52_{\pm 9.23}$	$10.40_{\pm 7.40}$	$14.96_{\pm 8.29}$
Retina	$13.73_{\pm 8.22}$	$13.96_{\pm 8.25}$	$31.55_{\pm 14.10}$	$27.84_{\pm 13.86}$	$16.14_{\pm 9.26}$
Breast	$12.60_{\pm 7.84}$	$23.16_{\pm 11.43}$	$18.46_{\pm 10.14}$	$39.06_{\pm 17.18}$	$43.00_{\pm 18.46}$
Blood	$11.55_{\pm 9.47}$	$9.37_{\pm 8.53}$	$33.78_{\pm 16.95}$	$11.14_{\pm 7.59}$	$12.04_{\pm 8.49}$
Tissue	$9.09_{\pm 5.21}$	$7.70_{\pm 7.43}$	$8.86_{\pm 8.36}$	$9.09_{\pm 7.66}$	$12.57_{\pm 7.26}$
OrganA	$10.14_{\pm 8.08}$	$6.24_{\pm 7.50}$	$7.17_{\pm 6.53}$	$5.62_{\pm 8.09}$	$7.97_{\pm 7.56}$
OrganC	$6.12_{\pm 5.92}$	$5.54_{\pm 7.73}$	$7.15_{\pm 6.25}$	$6.22_{\pm 6.22}$	$6.81_{\pm 7.00}$
OrganS	$5.68_{\pm 7.00}$	$6.60_{\pm 6.37}$	$7.01_{\pm 6.38}$	$6.02_{\pm 6.12}$	$8.73_{\pm 6.26}$

We further examine the impact of supervised IMAGENET1K initialization on model generalizability. To evaluate the overall impact of IMAGENET1K initialization compared to random initialization, we conducted paired t-tests for each test dataset across all cross-dataset training combinations including the in-domain setting where test and train splits come from the same dataset. The results demonstrate statistically significant improvements $(p < 0.05)$ across all datasets, with PneumoniaMNIST showing the most modest relative gain $(1.88 \pm 3.28\%)$ and OCTMNIST exhibiting the largest improvement $(13.22 \pm 25.59\%)$ when we switch from random weights to IMAGENET1K initialization. Detailed results are in Appendix D.3. While previous studies have raised concerns about the efficacy of transfer learning from natural to medical images due to significant distributional differences [67], our findings show that continual pre-training [35, 73] improves both in-domain performance and generalization.

ResNet-50 outperformed ViT-Small in most in-domain settings (8/11 datasets with random and 9/11 with IMAGENET1K initialization). However, as Figure 7 demonstrates, the performance gap between these architectures narrows considerably when assessing generalization rather than in-domain performance. Notably, with IMAGENET1K initialization, ViT-S actually surpasses ResNet-50 on 8/11 datasets, reversing the previous trend. This suggests that ViT-S benefits more from transfer learning when applied to new datasets.

Finally, we evaluate generalizability of representations learned through multi-domain training. Figure 6 shows that cross-dataset performance of models trained on Organ{A,C,S} is better than those trained on their individual constituents, suggesting that multi-domain training within a single modality may act as a regularizer. However, this benefit does not hold for mixed-modality combinations (Organ{A,S}PnePath), showing that multi-domain benefits are modality-dependent (Appendix D.3).

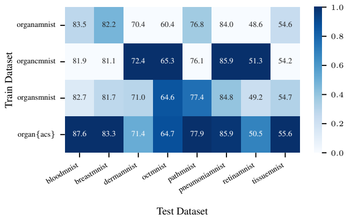


Figure 6: Normalized cross-dataset evaluation matrix showing averaged accuracies over eight SSL methods.

5 Discussion & Conclusion

We have conducted a comprehensive evaluation of popular SSL methods in medical imaging, addressing a crucial gap in understanding how these methods perform across diverse medical tasks. Based on our results and analysis, we offer the following recommendations for practitioners:

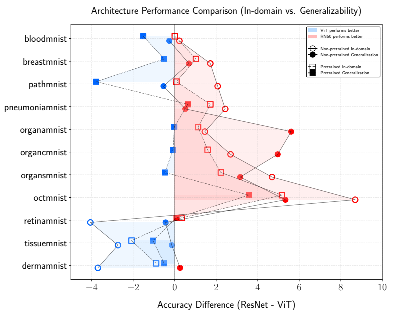


Figure 7: Performance differential between ResNet-50 and ViT-S (averaged over methods) for both ID and cross-dataset evaluations using different initializations.

Which self-supervised method to choose? MoCo v3 demonstrates remarkable versatility in medical imaging tasks when trained from scratch, achieving superior performance in 5/11 datasets for in-domain tasks and maintaining minimal accuracy drops in 4/11 datasets for cross-domain generalization. It also shows competitive performance in OOD detection. However, the effectiveness of certain SSL methods is significantly influenced by initialization strategy. Notably, DINO and BYOL transform from being among the lowest performers to achieving competitive in-domain results when initialized with IMAGENET1K weights, with DINO particularly excelling with ViT.

Should we start self-supervised pre-training with supervised IMAGENET1K weights? Self-supervised pretraining on medical images, initialized with IMAGENET1K weights, significantly outperforms random initialization for in-domain downstream tasks, mirroring established benefits previously observed in supervised learning setting [74]. Despite concerns that supervised pre-training might yield less general representations [54, 68], IMAGENET1K initialization enhances both in-domain and cross-dataset performance. However, this advantage does not consistently extend to OOD detection, as IMAGENET1K initialization degrades performance for certain backbone and model combinations. Overall, supervised and self-supervised IMAGENET1K initialization offer comparable performance for in-domain accuracy and OOD detection with no clear advantage for either.

Which model architecture is better? Choosing between these architectures involves key trade-offs: While ResNet-50 excels in in-domain classification (outperforming ViT-Small in 8/11 datasets when trained from scratch and 9/11 with IMAGENET1K initialization), ViT-Small demonstrates superior performance in OOD detection. Interestingly, when trained with IMAGENET1K initialization and evaluated on generalization tasks, ViT-Small reverses the in-domain trend, surpassing ResNet-50 in 8/11 datasets. However, ViT shows greater sensitivity to label scarcity, with performance dropping more steeply when training data is limited. IMAGENET1K initialization reduces the performance gap between architectures, though these fundamental trade-offs remain.

Is multi-domain SSL effective? Dataset composition plays a crucial role in model performance. Models trained on heterogeneous datasets (Organ{A,S}PnePath) demonstrated superior OOD detection, achieving higher AUROC scores than those trained on homogeneous (Organ{A,C,S}) or single-domain datasets. While Organ{A,C,S} improves in-domain accuracy and generalization over individual Organ datasets, this pattern does not hold for the more diverse Organ{A,S}PnePath.

Extended Evaluation of Key Findings: We conducted a series of short experiments using bigger architectures, higher resolution images, and an additional dataset AIROGS [22]. The results (in Appendix G) confirm that our key findings consistently hold across these settings.

6 Acknowledgements

B.K. gratefully acknowledges partial support from the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) Master's Scholarships. The work described in this paper was conducted in the framework of the Graduate School 2543/1 "Intraoperative Multi-Sensory Tissue Differentiation in Oncology" (project ID 40947457) funded by the German Research Foundation (DFG - Deutsche Forschungsgemeinschaft). This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) – EXC number 2064/1 – Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Valay Bundele and Zohreh Ghaderi.

References

[1] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020. ISSN 2352-3409.

[2] Muhammad Muneeb Afzal, Muhammad Osama Khan, and Yi Fang. A comprehensive benchmark of supervised and self-supervised pre-training on multi-view chest x-ray classification. In *Medical Imaging with Deep Learning*, 2024. URL https://openreview.net/forum?id=YUMVjxdIqn.

[3] Muhammad Muneeb Afzal, Muhammad Osama Khan, and Yi Fang. A comprehensive benchmark of supervised and self-supervised pre-training on multi-view chest x-ray classification. *Medical Imaging with Deep Learning*, 1, 2024.

[4] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021. URL https://arxiv.org/abs/2104.11178.

[5] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021. URL https://arxiv.org/abs/2103.15691.

[7] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification, 2021. URL https://arxiv.org/abs/2101.05224.

[8] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S. Sara Mahdavi, Ellery Wulczyn, Boris Babenko, Megan Wilson, Aaron Loh, Po-Hsuan Cameron Chen, Yuan Liu, Pinal Bavishi, Scott Mayer McKinney, Jim Winkens, Abhijit Guha Roy, Zach Beaver, Fiona Ryan, Justin Krogue, Mozziyar Etemadi, Umesh Telang, Yun Liu, Lily Peng, Greg S. Corrado, Dale R. Webster, David Fleet, Geoffrey Hinton, Neil Houlsby, Alan Karthikesalingam, Mohammad Norouzi, and Vivek Natarajan. Robust and efficient medical imaging with self-supervision, 2022. URL https://arxiv.org/abs/2205.09723.

[9] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. URL https://arxiv.org/abs/2105.04906.

[10] Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. Confidence-based out-of-distribution detection: A comparative study and analysis, 2021. URL https://arxiv.org/abs/2107.02568.

[11] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. ISSN 1361-8415.

[12] Behzad Bozorgtabar, Guillaume Vray, Dwarikanath Mahapatra, and Jean-Philippe Thiran. Sood: Self-supervised out-of-distribution detection under domain shift for multi-class colorectal cancer tissue types. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3317–3326, 2021. doi: 10.1109/ICCVW54120.2021.00371.

[13] Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. Medianomaly: A comparative study of anomaly detection in medical images, 2024. URL https://arxiv.org/abs/2404.04518.

[14] Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. Medianomaly: A comparative study of anomaly detection in medical images. *Medical Image Analysis*, page 103500, 2025.

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

[17] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.02057.

[18] Yunfeng Chen, Yalan Lin, Xiaodie Xu, Jinzhen Ding, Chuzhao Li, Yiming Zeng, Weifang Xie, and Jianlong Huang. Multi-domain medical image translation generation for lung image classification based on generative adversarial networks. *Computer Methods and Programs in Biomedicine*, 229:107200, 2023. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2022.107200. URL https://www.sciencedirect.com/science/article/pii/S0169260722005818.

[19] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 168–172. IEEE, 2019.

[20] MMSelfSup Contributors. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. https://github.com/open-mmlab/mmselfsup, 2021.

[21] Victor G. Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. Solo-learn: A library of self-supervised methods for visual representation learning, 2022. URL https://arxiv.org/abs/2108.01775.

[22] Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, Devika R G, Hrishikesh P S, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikan, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. Airogs: Artificial intelligence for robust glaucoma screening challenge, 2023. URL https://arxiv.org/abs/2302.01738.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[24] Sebastian Doerrich, Francesco Di Salvo, Julius Brockmann, and Christian Ledig. Rethinking model prototyping through the medmnist+ dataset collection, 2024. URL https://arxiv.org/abs/2404.15786.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[26] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations, 2021. URL https://arxiv.org/abs/2104.14548.

[27] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, May 2022. ISSN 1558-0792. doi: 10.1109/msp.2021.3134634. URL http://dx.doi.org/10.1109/MSP.2021.3134634.

[28] Alex Fedorov, Eloy Geenjaar, Lei Wu, Thomas P. DeRamus, Vince D. Calhoun, and Sergey M. Plis. Tasting the cake: evaluating self-supervised generalization on out-of-distribution multimodal mri data, 2022. URL https://arxiv.org/abs/2103.15914.

[29] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet, 2023. URL https://arxiv.org/abs/2302.11893.

[30] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning, 2019. URL https://arxiv.org/abs/1905.01235.

[31] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

[32] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

[33] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2024. doi: 10.1109/TPAMI.2024.3415112.

[34] Charles Guille-Escuret, Pau Rodriguez, David Vazquez, Ioannis Mitliagkas, and Joao Monteiro. Cadet: Fully self-supervised out-of-distribution detection with contrastive learning, 2023. URL https://arxiv.org/abs/2210.01742.

[35] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks, 2020. URL https://arxiv.org/abs/2004.10964.

[36] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20824–20834, 2022.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. URL https://arxiv.org/abs/1911.05722.

[39] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. 2016.

[40] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty, 2019. URL https://arxiv.org/abs/1906.12340.

[41] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3*, pages 3–13. Springer, 2021.

[42] S.C. Huang, A. Pareek, M. Jensen, et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6 (74), 2023. doi: 10.1038/s41746-023-00811-0. URL https://doi.org/10.1038/s41746-023-00811-0.

[43] Zhe Huang, Ruijie Jiang, Shuchin Aeron, and Michael C. Hughes. Systematic comparison of semi-supervised and self-supervised learning for medical image classification, 2024. URL https://arxiv.org/abs/2307.08919.

[44] Marie Humbert-Droz, Pritam Mukherjee, and Olivier Gevaert. Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: Case study for the extraction of symptoms from clinical notes. *JMIR Medical Informatics*, 10(3):e32903, 2022. doi: 10.2196/32903.

[45] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets, 2023. URL https://arxiv.org/abs/2212.04690.

[46] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16 (e1002730), 2019. ISSN 1549-1676.

[47] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172:1122–1131.e9, 2018. ISSN 0092-8674.

[48] Umar Khalid, Ashkan Esmaeili, Nazmul Karim, and Nazanin Rahnavard. Rodd: A self-supervised approach for robust out-of-distribution detection, 2022. URL https://arxiv.org/abs/2204.02553.

[49] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning, 2019. URL https://arxiv.org/abs/1901.09005.

[50] Sonja Kunzmann, Mathias Öttl, Prathmesh Madhu, Felix Denzinger, and Andreas Maier. An unobtrusive quality supervision approach for medical image annotation, 2022. URL https://arxiv.org/abs/2211.06146.

[51] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018. URL https://arxiv.org/abs/1807.03888.

[52] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020.

[53] Jingyao Li, Pengguang Chen, Shaozuo Yu, Zexin He, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need, 2023. URL https://arxiv.org/abs/2302.02615.

[54] Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance, 2022. URL https://arxiv.org/abs/2110.05025.

[55] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, Adrian Galdran, J. M. Poorneshwaran, Hao Liu, Jie Wang, Yerui Chen, Prasanna Porwal, Gavin Siew Wei Tan, Xiaokang Yang, Chao Dai, Haitao Song, Mingang Chen, Huating Li, Weiping Jia, Dinggang Shen, Bin Sheng, and Ping Zhang. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3:100512, 2022. ISSN 2666-3899.

[56] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection, 2020.

[57] Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7):637, 2012. ISSN 1548-7105.

[58] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[59] Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification, 2024. URL https://arxiv.org/abs/2407.12210.

[60] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images?, 2021. URL https://arxiv.org/abs/2108.09038.

[61] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. What makes transfer learning work for medical images: Feature reuse & other factors, 2022. URL https://arxiv.org/abs/2203.01825.

[62] Sina Mohseni, Mandar Pitale, Jbs Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:212803476.

[63] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in Biology and Medicine*, 128:104115, January 2021. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2020.104115. URL http://dx.doi.org/10.1016/j.compbiomed.2020.104115.

[64] Vivek Narayanaswamy, Yamen Mubarka, Rushil Anirudh, Deepta Rajan, and Jayaraman J. Thiagarajan. Exploring inlier and outlier specification for improved medical ood detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4589–4598, October 2023.

[65] Ifeoma Veronica Nwabufo, Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. Self-supervised visualisation of medical image datasets. *arXiv preprint arXiv:2402.14566*, 2024.

[66] Ece Ozkan and Xavier Boix. Multi-domain improves out-of-distribution and data-limited scenarios for medical image analysis, 2024. URL https://arxiv.org/abs/2310.06737.

[67] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging, 2019. URL https://arxiv.org/abs/1902.07208.

[68] Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, and Diane Larlus. No reason for no supervision: Improved generalization in supervised models, 2023. URL https://arxiv.org/abs/2206.15369.

[69] Marc-Andre Schulz, Bertrand Thirion, Alexandre Gramfort, Gaël Varoquaux, and Danilo Bzdok. Label scarcity in biomedicine: Data-rich latent factor discovery enhances phenotype prediction, 2021. URL https://arxiv.org/abs/2110.06135.

[70] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[71] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf.

[72] Igor Susmelj, Matthias Heller, Philipp Wirth, Jeremy Prescott, Malte Ebner, and et al. Lightly. URL https://github.com/lightly-ai/lightly.

[73] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B. Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis, 2021. URL https://arxiv.org/abs/2108.05930.

[74] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, May 2016. ISSN 1558-254X. doi: 10.1109/tmi.2016.2535302. URL http://dx.doi.org/10.1109/TMI.2016.2535302.

[75] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5: 1–9, 2018. ISSN 2052-4463.

[76] Stefano Woerner, Arthur Jaques, and Christian F. Baumgartner. A comprehensive and easy-to-use multi-domain multi-task medical imaging meta-dataset (medimeta), 2024. URL https://arxiv.org/abs/2404.16000.

[77] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification, 2022. URL https://arxiv.org/abs/2210.12843.

[78] Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. Efficient multiple organ localization in ct images using 3d region proposal network. *IEEE Transactions on Medical Imaging*, 38(8):1885–1898, 2019.

[79] Siyuan Yan, Zhen Yu, Chi Liu, Lie Ju, Dwarikanath Mahapatra, Brigid Betz-Stablein, Victoria Mar, Monika Janda, Peter Soyer, and Zongyuan Ge. Prompt-driven latent domain generalization for medical image classification. *IEEE Transactions on Medical Imaging*, 2024.

[80] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1), January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01721-8. URL http://dx.doi.org/10.1038/s41597-022-01721-8.

[81] Karina Zadorozhny, Patrick Thoral, Paul Elbers, and Giovanni Cinà. Out-of-distribution detection for medical applications: Guidelines for practical evaluation, 2022. URL https://arxiv.org/abs/2109.14885.

[82] Muhammad Zaid, Shafaqat Ali, Mohsen Ali, Sarfaraz Hussein, Asma Saadia, and Waqas Sultani. Identifying out of distribution samples for skin cancer and malaria images. *Biomedical Signal Processing and Control*, 78:103882, 2022. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2022.103882. URL https://www.sciencedirect.com/science/article/pii/S1746809422003937.

[83] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. URL https://arxiv.org/abs/2103.03230.

[84] Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts, 2021.

[85] Yu Zhang, Jing Chen, Xiangxun Ma, Gang Wang, Uzair Aslam Bhatti, and Mengxing Huang. Interactive medical image annotation using improved attention u-net with compound geodesic distance. *Expert Syst. Appl.*, 237(PA), February 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.121282. URL https://doi.org/10.1016/j.eswa.2023.121282.

[86] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation, 2021. URL https://arxiv.org/abs/2107.09282.

[87] Haoran Zhu, Boyuan Chen, and Carter Yang. Understanding why vit trains badly on small datasets: An intuitive perspective, 2023. URL https://arxiv.org/abs/2302.03751.