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

Valay Bundele*, Karahan Sarıtaş*, Bora Kargi*, Oğuz Ata Çal*, Kıvanç Tezören*, Zohreh Ghaderi, Hendrik Lensch

(*Equal contribution)

University of Tübingen

Funded by Deutsche Forschungsgemeinschaft German Research Foundation

BMVC 2025

imprs-is EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Introduction

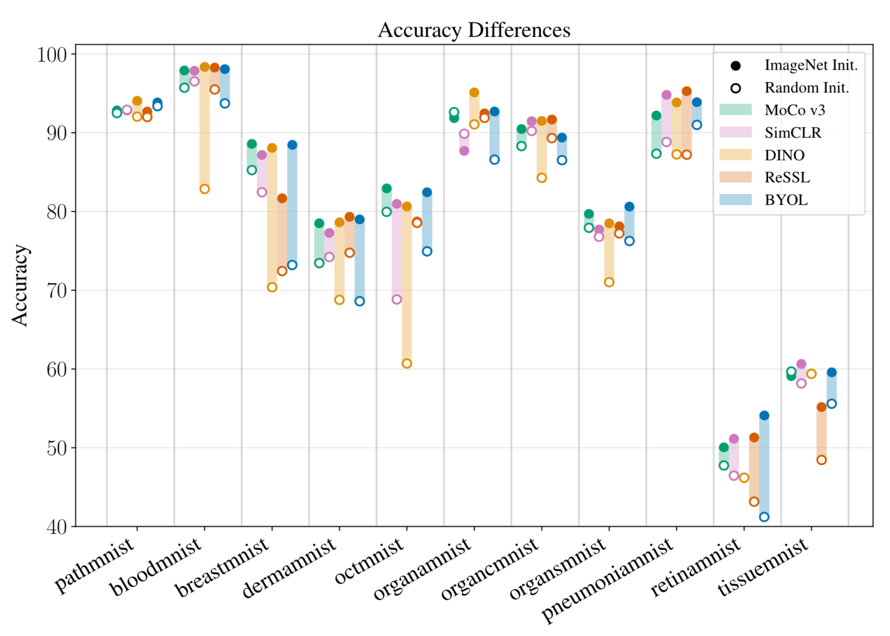
- While self-supervised learning (SSL) has shown impressive results in medical imaging, existing studies are often limited in scope:
  - focusing on specific datasets or modalities.
  - evaluating only isolated aspects of model performance.
- Poses a significant challenge, as models deployed in critical medical settings must also show robust performance and generalizability across diverse datasets and varying conditions.
- Conduct a rigorous investigation of SSL for medical imaging, evaluating 8 major SSL methods across 11 real-world medical datasets, spanning three core dimensions:
  - in-domain performance under varying label proportions (1%, 10%, and 100%).
  - cross-dataset generalization.
  - 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.

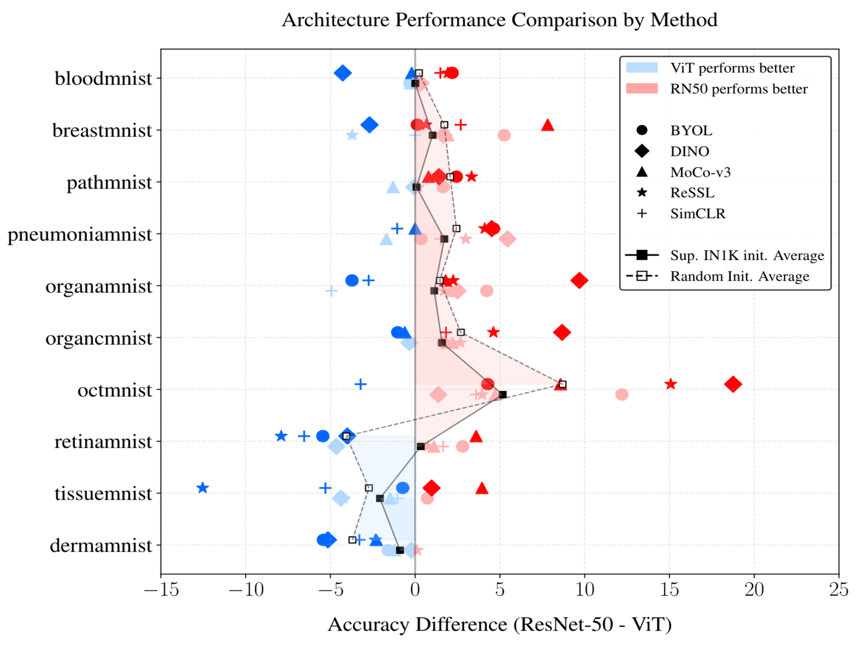
Methodology

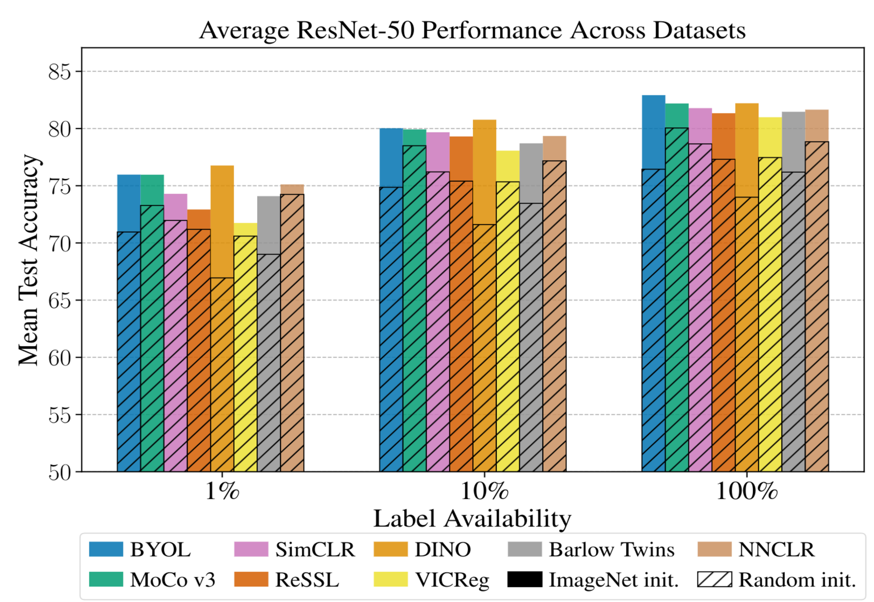
- SSL methods: SimCLR, DINO, BYOL, ReSSL, MoCo v3, NNCLR, VICREG, and Barlow Twins.
- Eleven medical datasets utilized in this study are derived from MedMNIST.
- Five different pre-training schemes:
  - supervised training with (1) random initialization, (2) supervised IMAGENET1K init
  - self-supervised pre-training with (3) random initialization, (4) supervised IMAGENET1K, and (5) self-supervised IMAGENET1K initializations.
- Generalizability is assessed by conducting cross-dataset experiments using the dataset collection.
- OOD detection is done using Mahalanobis distance.
- For multi-domain learning, we use two combinations, namely,
  - Organ{A,C,S}: OrganAMNIST, OrganCMNIST and OrganSMNIST datasets.
  - Organ{A,S}PnePath: OrganAMNIST (CT), OrganSMNIST (CT), PathMNIST (Colon Pathology), and PneumoniaMNIST (Chest X-Ray).

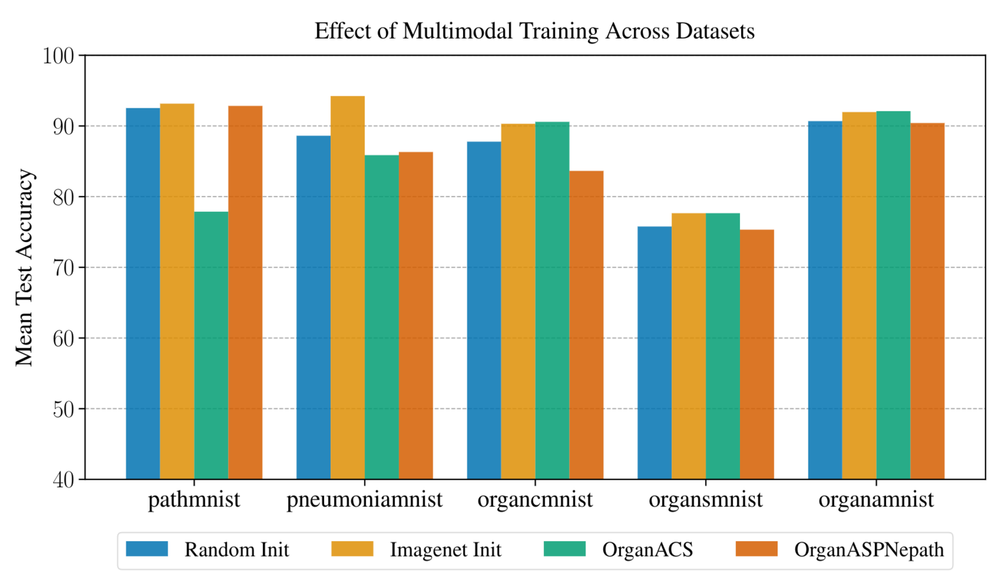
In-Domain Performance

- Evaluate in-domain performance by pre-training ResNet-50 or ViT-small with each SSL method.
- MoCo v3 achieves the highest accuracy on 5 of 11 datasets, followed closely by BYOL, SimCLR.
- SSL outperforms supervised learning on 7 of 11 datasets when both use random initialization.



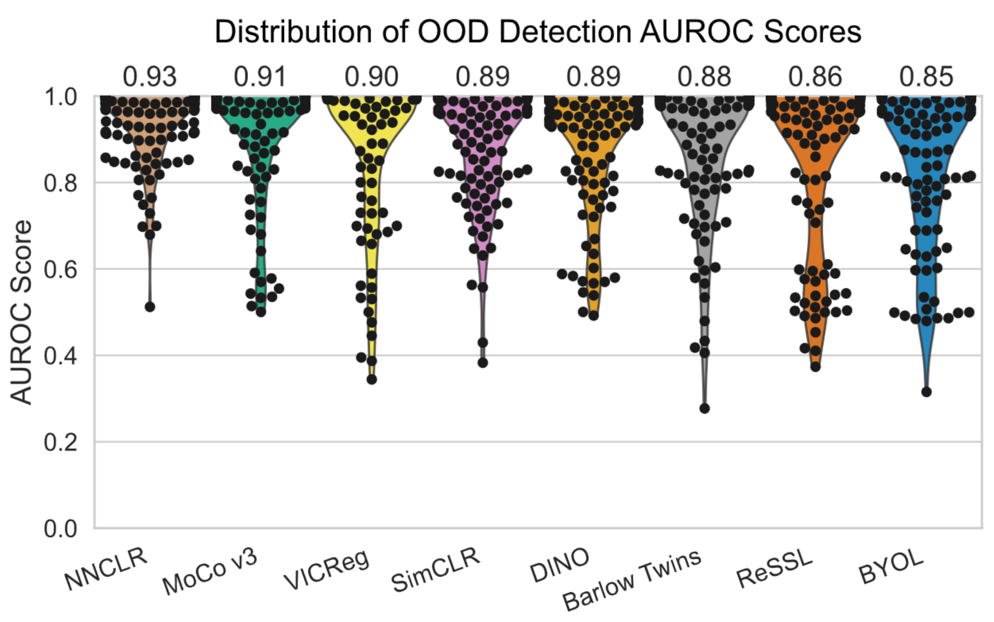


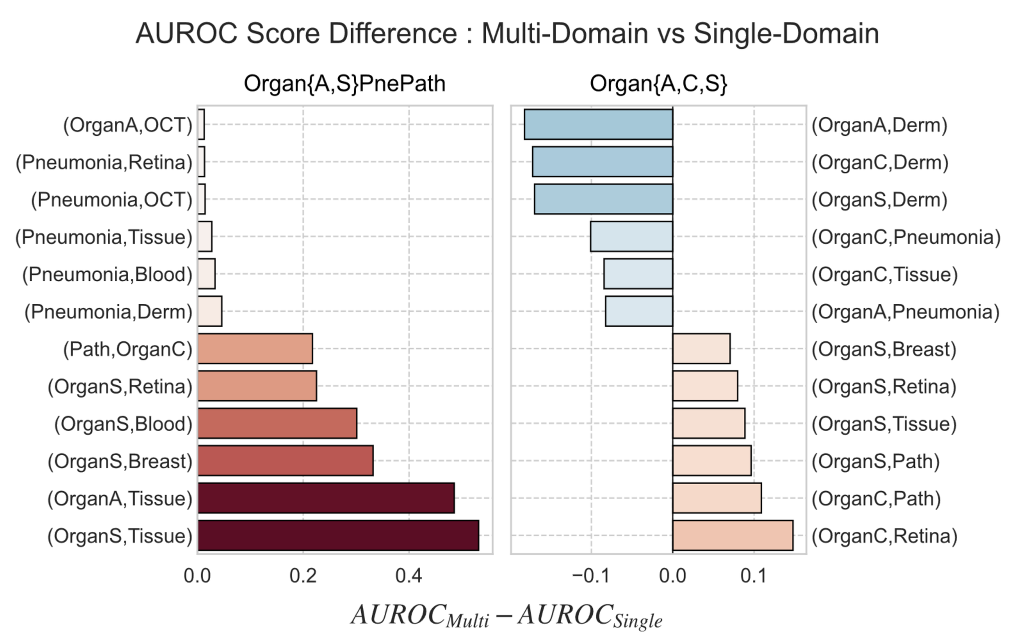


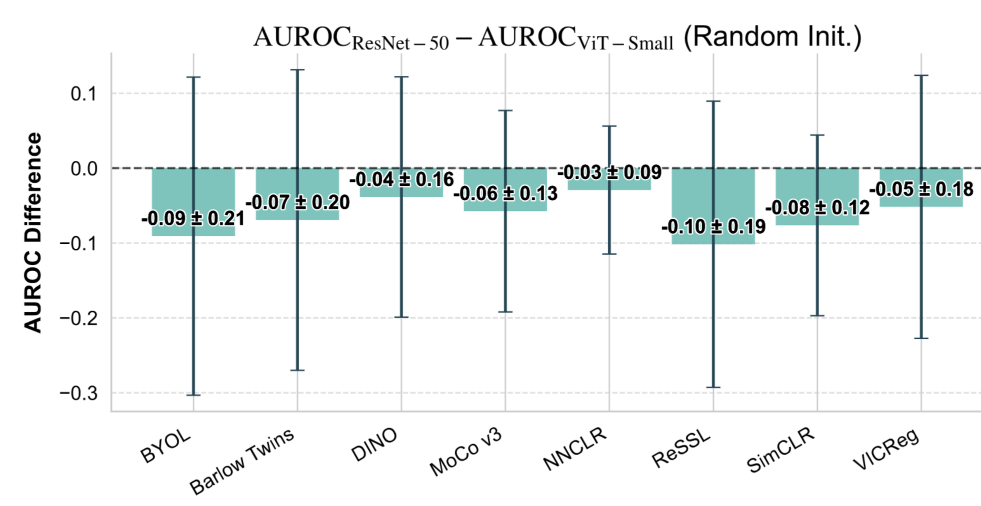


Out-of-Distribution Detection

- Evaluate OOD 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 ID dataset and assessed to identify OOD samples from other datasets.
- NNCLR shows the highest OOD detection performance with ResNet-50 while MoCo v3 shows highest performance with ViT-small.







Generalizability

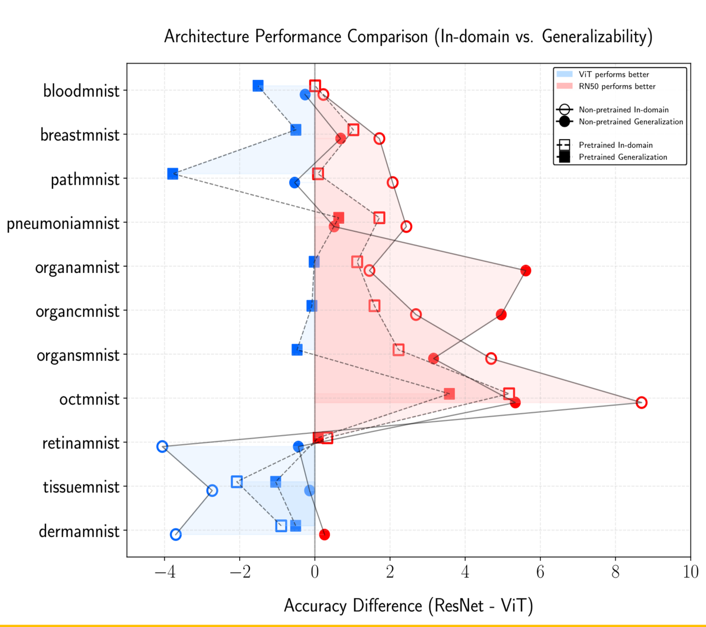
- Assess transferability of SSL representations via cross-dataset evaluations.
- Specifically, we first pre-train the encoder on one dataset (source) and then train a linear classifier on each of the remaining datasets individually (target), using a frozen encoder.
- Generalizability is measured by relative accuracy drop:

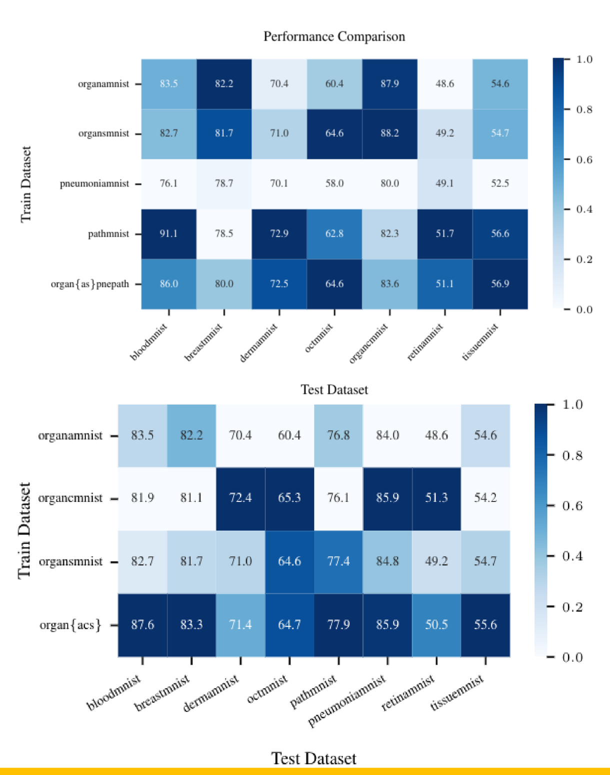
$$\Delta ACC = \frac{ACC_{ID} - ACC_{transfer}}{ACC_{ID}} \times 100$$

- For each source dataset, we report average accuracy drop, revealing how well learned representations generalize and which datasets yield the most transferable features.
- ACC_{ID} represents accuracy when training and testing are performed on the same dataset, and $ACC_{transfer}$ denotes accuracy when model is pre-trained on a different dataset.
- To evaluate impact of IMAGENET1K init, we conducted paired t-tests for each test dataset across all cross-dataset training combinations including the in-domain setting.
- SimCLR, MoCo v3 excel in generalization while models pre-trained on Path or Organ{C,S}MNIST show lowest accuracy drops, suggesting they offer most transferable representations.

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	$5.68_{\pm 7.00}$	$6.60_{\pm 6.37}$	$7.01_{\pm 6.38}$	$6.02_{\pm 6.12}$	$8.73_{\pm 6.26}$





Dataset	Random Init.	IMAGENET1K Init.	Improvement	p-value
BloodMNIST	80.52	88.41	$13.52 \pm 27.58\%$	0.0000
BreastMNIST	79.68	82.26	$3.47 \pm 6.47\%$	0.0000
DermaMNIST	70.67	73.26	$3.76 \pm 4.48\%$	0.0000
OctMNIST	58.33	68.24	$22.08 \pm 31.16\%$	0.0000
OrganAMNIST	81.54	88.94	$12.79 \pm 26.96\%$	0.0000
OrganCMNIST	79.57	85.15	$10.86 \pm 27.45\%$	0.0002
OrganCMNIST	66.31	72.30	$13.55 \pm 29.62\%$	0.0000
PathMNIST	72.90	81.86	$16.95 \pm 32.10\%$	0.0000
PneumoniaMNIST	83.50	86.52	$3.93 \pm 7.28\%$	0.0000
RetinaMNIST	48.82	50.63	$4.00 \pm 7.56\%$	0.0000
TissueMNIST	52.99	57.46	$9.50 \pm 13.30\%$	0.0000

Comparison of random initialization versus IMAGENET1K initialization across different medical datasets. Results show mean accuracy (%) for both initialization methods, relative improvement, and statistical significance.

Conclusion

- MoCov3 performs best on 5/11 in-domain datasets and shows minimal drops on 4/11 cross-domain datasets. It also performs competitively in OOD detection.
- Medical SSL pretraining (initialized from ImageNet1K) outperforms random initialization for in-domain and cross-domain tasks but not consistently for OOD detection.
- While ResNet-50 excels in in-domain tasks, ViT-Small shows superior performance in OOD detection. When trained with IMAGENET1K init, ViT-Small outperforms ResNet-50 in most cross-domain tasks.
- Models trained on heterogeneous datasets (Organ{A,S}PnePath) show better OOD detection; Organ{A,C,S} improves in-domain and generalization, but this trend breaks for Organ{A,S}PnePath.