scientific reports

Check for updates

OPEN

Rethinking model prototyping through the MedMNIST+ dataset collection

Sebastian Doerrich[1]✉, Francesco Di Salvo[1], Julius Brockmann[1,2] & Christian Ledig[1]

The integration of deep learning based systems in clinical practice is often impeded by challenges rooted in limited and heterogeneous medical datasets. In addition, the field has increasingly prioritized marginal performance gains on a few, narrowly scoped benchmarks over clinical applicability, slowing down meaningful algorithmic progress. This trend often results in excessive fine-tuning of existing methods on selected datasets rather than fostering clinically relevant innovations. In response, this work introduces a comprehensive benchmark for the MedMNIST+ dataset collection, designed to diversify the evaluation landscape across several imaging modalities, anatomical regions, classification tasks and sample sizes. We systematically reassess commonly used Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) architectures across distinct medical datasets, training methodologies, and input resolutions to validate and refine existing assumptions about model effectiveness and development. Our findings suggest that computationally efficient training schemes and modern foundation models offer viable alternatives to costly end-to-end training. Additionally, we observe that higher image resolutions do not consistently improve performance beyond a certain threshold. This highlights the potential benefits of using lower resolutions, particularly in prototyping stages, to reduce computational demands without sacrificing accuracy. Notably, our analysis reaffirms the competitiveness of CNNs compared to ViTs, emphasizing the importance of comprehending the intrinsic capabilities of different architectures. Finally, by establishing a standardized evaluation framework, we aim to enhance transparency, reproducibility, and comparability within the MedMNIST+ dataset collection as well as future research. Code is available at (https://github.com/sdoerrich97/rethinking-model-prototyping-MedMNISTPlus).

Keywords Benchmarking, Prototyping recommendations, Medical image classification, Standardized evaluation framework, Foundation models

In recent years, significant strides in deep learning (DL) have reshaped various domains, from image classification to natural language processing[1]. This progress was driven by the development of increasingly sophisticated models, exemplified by architectures like the Transformer[2] for text or Vision Transformer (ViT)[3] for images. Moreover, advanced training methodologies, including self-supervised contrastive methods such as CLIP[4] for image and text pairs, and DINO[5,6] for image pairs, have enabled the training of complex models without the need for exhaustive labeling efforts. Simultaneously, there is an accumulating interest in integrating machine learning techniques into medical imaging, where DL models are approaching performance comparable to medical experts on specific tasks[7] and software applications are beginning to receive clinical certifications[8]. Despite this progress and the exponential growth of DL-related publications across various medical fields in the past few years[9], the adoption of DL algorithms in daily clinical practice has been comparatively slow[10].

One major obstacle is the scarcity of appropriate datasets, often characterized by limited sample sizes and heterogeneous image acquisition conditions[11–13], thereby posing challenges to the generalizability of supervised DL algorithms. Ongoing progress in domain adaptation (DA) and domain generalization (DG) techniques aims to increase algorithmic robustness through aligning feature distributions[14] or acquiring domain-invariant features[15]. However, the generalizability of these methods across diverse domains remains a significant challenge, constraining their real-world applicability[16].

In addition, there is a concerning trend in DL research towards prioritizing the adaptation and scaling of existing methodologies in order to achieve incremental performance improvements on influential benchmarks[17] rather than addressing clinically relevant needs[18]. This trend is particularly pronounced in academic research,

[1]University of Bamberg, xAILab Bamberg, Bamberg 96047, Germany. [2]Ludwig Maximilian University of Munich, Munich 80539, Germany. ✉email: sebastian.doerrich@uni-bamberg.de

where the incentive structure often prioritizes quantity over relevance, leading to the incorporation of additional complexity into existing methodologies often at the expense of increased computational requirements[19]. While benchmark datasets play a crucial role in coordinating machine learning research and facilitating standardized evaluations[20], overreliance on a handful of influential yet narrowly scoped benchmarks may stifle innovation and exacerbate inherent biases within these datasets such as the underrepresentation of certain demographic groups[21,22]. The latter, in particular, limits the applicability of current DL techniques across diverse patient populations, thereby impeding their real-world deployment[23]. Instead, research endeavors should focus more on proposing new benchmarks to diversify the landscape, mitigate bias-induced challenges, and cover a broader range of real-world tasks. Rather than solely determining a winner based on state-of-the-art performance, benchmarking should promote understanding to drive impactful algorithmic development and alternative evaluation methods[17].

Furthermore, the limitations of scaling alone are becoming increasingly evident, as larger models start to falter in model trustworthiness[24,25] or performance on well-specified tasks[26]. Nonetheless, there is a paramount trend of increasing hardware and compute requirements estimated via the total number of FLOPs, and the number of trainable parameters in deep learning architectures[27]. This further impedes the application of these approaches in the clinical environment. Therefore, it is imperative to explore qualitative enhancements alongside quantitative scaling in DL research as called for by A. Goyal and Y. Bengio[28], particularly in the context of real-world medical applications.

Research endeavors should prioritize the creation of larger and more diverse datasets and benchmarks, with a focus on incorporating additional inductive biases and fostering the continuous development of more sophisticated approaches. The recent emergence of foundation models exemplifies this direction. These models, pre-trained on extensive datasets, offer the potential to enhance performance by capturing intricate patterns and serving as a foundational basis for further fine-tuning[29]. Existing works, building on top of these models, can be readily evaluated across diverse benchmarks due to their high transferability to new datasets and tasks, as well as their remarkable zero- or few-shot performance[30,31]. This facilitates a more comprehensive assessment of these methods without necessitating extensive retraining.

In this work, we aim to contribute to this effort by reassessing traditional DL models and training schemes, and presenting a new benchmark in the context of medical image classification. Our objective is to reevaluate commonly held assumptions regarding these methodologies, thereby enhancing and confirming comprehension of their inherent strengths and limitations as well as diversifying the benchmark landscape. Consequently, we will offer recommendations and insights for prototyping, model development, and deployment. To this end, we extend upon the existing MedMNIST v2 classification benchmark[32], using the recently introduced MedMNIST+ database[33]. MedMNIST v2 offers a collection of 12 distinct biomedical 2D datasets, ranging from Chest-X-ray to Dermatology, in a MNIST-like[34] resolution of 28×28 pixels for medical image analysis. Its limitation to 28×28 images represented a critical constraint for comprehensive method evaluation. However, this has been addressed with the introduction of MedMNIST+, extending the previous dataset collection with resolutions: 64×64, 128×128, and 224×224 pixels. By systematically benchmarking a diverse array of baseline models and training paradigms, including selective convolutional and Transformer-based models using both end-to-end training and linear probing on this distinct multi-dimensional database, our goal is to provide critical insights into the strengths and weaknesses of these techniques. Furthermore, we investigate the integration of the k-nearest neighbors (k-NN) classifier into the feature space of these models, aiming to enhance computational efficiency and interpretability. Given that most clinically validated and regulated systems currently rely on supervised deep learning, in particular CNN-based architectures[35,36], we prioritize these widely adopted models in our analysis. Our primary aim is to re-investigate whether compute-intensive architectures are always necessary, whether higher resolution input consistently improves model performance, and whether end-to-end training is always optimal. Additionally, we want to foster greater transparency, reproducibility, and comparability in future research endeavors within the domain of medical image analysis. Key contributions of our work include:

- Systematic benchmarking of a wide range of commonly used models across diverse medical datasets, accounting for variations in resolutions, tasks, sample sizes, and class distributions.
- Identification of systematic strengths and weaknesses inherent in traditional models within the context of medical image classification.
- Reevaluation of prevalent assumptions with respect to model design, training schemes and input resolution requirements.
- Presentation of a solid baseline performance for MedMNIST+ and a standardized evaluation framework for assessing future model performance in medical image classification.
- Formulation of recommendations and take-aways for model development and deployment.

Experiments and results

Datasets

The dataset selection employed in this work originates from MedMNIST v2[32], initially introduced at a resolution of 28×28 pixels and recently expanded by MedMNIST+[33] to four distinct image resolutions, namely 28×28 , 64×64, 128×128, and 224×224. The collection comprises twelve 2D datasets that are curated from carefully selected sources, encompassing primary data modalities such as X-ray, OCT, ultrasound, CT, and electron microscope. Furthermore, these datasets cater to diverse classification tasks, including binary/multi-class, ordinal regression, and multi-label classification, spanning a wide range of dataset scales, ranging from 780 samples for *Breast* up to 236,386 for *Tissue*. The details of each dataset including data source, imaging modality, type of classification task (along with the number of classes), and the publicly available data splits, provided by MedMNIST[33] and forming a one-to-one correspondence with our benchmark, are described in Table 1.

Dataset	Source	Imaging Modality	Task (# Classes)	Number of Samples Train / Val / Test	Class Imbalance Train / Val / Test
Blood	A. Acevedo et al.[37]	Blood Cell Microscope	MC (8)	11, 959 / 1, 712 / 3, 421	0.96 / 0.96 / 0.96
Breast	W. Al-Dhabyani et al.[38]	Breast Ultrasound	BC (2)	546 / 78 / 156	0.84 / 0.84 / 0.84
Chest	X. Wang et al.[39]	Chest X-Ray	ML-BC (2)	78,468 / 11,219 / 22,433	0.26 / 0.26 / 0.26
Derma	P. Tschandl et al.[40] N. Codella et al.[41]	Dermatoscope	MC (7)	7,007 / 1,003 / 2,005	0.58 / 0.58 / 0.58
OCT	D. S. Kermany et al.[42]	Retinal OCT	MC (4)	97,477 / 10,832 / 1,000	0.84 / 0.84 / 1.00
OrganA	P. Bilic et al.[43] X. Xu et al.[44]	Abdominal CT	MC (11)	34,561 / 6,491 / 17,778	0.96 / 0.95 / 0.96
OrganC	P. Bilic et al.[43] X. Xu et al.[44]	Abdominal CT	MC (11)	12,975 / 2,392 / 8,216	0.95 / 0.95 / 0.95
OrganS	P. Bilic et al.[43] X. Xu et al.[44]	Abdominal CT	MC (11)	13,932 / 2,452 / 8,827	0.93 / 0.96 / 0.94
Path	J. N. Kather et al.[45]	Colon Pathology	MC (11)	89,996 / 10,004 / 7,180	0.99 / 0.99 / 0.96
Pneumonia	D. S. Kermany et al.[42]	Chest X-Ray	BC (2)	4,708 / 524 / 624	0.82 / 0.82 / 0.95
Retina	R. Liu et al.[46]	Fundus Camera	OR (5)	1,080 / 120 / 400	0.87 / 0.86 / 0.87
Tissue	V. Ljosa et al.[47]	Kidney Cortex Microscope	MC (8)	165,466 / 23,640 / 47,280	0.87 / 0.87 / 0.87

Table 1. Dataset details including data source, imaging modality, type of classification task (with number of classes), predefined data splits, and class imbalance measured using Shannon's Equitability (ranging from 0 for total imbalance to 1 for perfect balance) for each split. (ML: Multi-Label, MC: Multi-Class, BC: Binary-Class, OR: Ordinary Regression).

Additionally, we report Shannon's Equitability for each split to quantify class imbalance. A visual comparison of all datasets across the four evaluated image resolutions is presented in Figure 1.

Benchmark evaluation

We report the performance assessment of each model selected from our diverse pool, following the methodology outlined in Section Training pipeline. Our evaluation encompasses all datasets, image resolutions, and training schemes described earlier. The summary of average accuracy (ACC) and area under the receiver operating characteristic curve (AUC) across all datasets is provided in Table 2. In addition, detailed performance evaluations for each dataset individually can be found in the appendix, spanning Supplementary Tables S1 to S12. To ensure the reliability of our assessments for all datasets, we report the mean and standard deviation of ACC and AUC for three random seeds.

It is important to note that the k-NN approach does not furnish an AUC score. This peculiarity arises from its distinct classification methodology, which involves a majority vote on the labels associated with the k-closest training embeddings (lowest cosine distance). Although the possibility exists to convert this voting into a probability distribution by normalizing the vote through a division by k, such a metric would lack reliability and accuracy as it fails to provide a comprehensive probability distribution across all classes, but solely among neighboring ones. Consequently, the AUC score would be significantly influenced by factors such as the choice of k, local density, and data imbalance, prompting us to exclude this in our evaluation.

As anticipated, end-to-end training yields the highest overall performance for all training schemes, and higher resolutions appear to enable all models across all training schemes to exhibit performance enhancements compared to lower resolutions. However, these enhancements begin to saturate when transitioning from inputs of 128×128 pixels to 224×224 pixels. Despite the increased data information, characterized by a fourfold increase in pixel count, all models and training schemes across all datasets show only marginal improvements or, in some cases, even worse overall performance. This correspondence is visually depicted in Figure 2, where the accuracy distributions of each model across all datasets are displayed for each training scheme and input resolution, respectively. Furthermore, the performance relationships of the models to each other for a specific training scheme and input resolution remain largely consistent. This observation challenges the common assumption that evaluations solely on higher image resolutions (*e.g.* above 200×200 pixels) are deemed valid, while evaluations on lower input resolutions are generally considered less meaningful, since the performance trends appear to be resolution independent. This in turn supports the utilization of lower resolution inputs, particularly during the prototyping phase of model development, as they generally allow for faster processing speeds while demanding fewer computational resources.

Moreover, our analysis reveals that more extensive self-supervised pretraining strategies such as CLIP and DINO, when compared to ImageNet pretraining, do not necessarily lead to improved performance for end-to-end trained models. However, they do demonstrate enhanced performance for linear probing and the integration of k-NN. Particularly notable is the performance of DINO, which achieves results close to the end-to-end trained baseline while requiring minimal training (linear probing) or no training at all (k-NN). This suggests that the latter two training schemes benefit from extensive pretraining even if conducted on unrelated images. This raises questions about whether more sophisticated foundation models can further narrow the performance

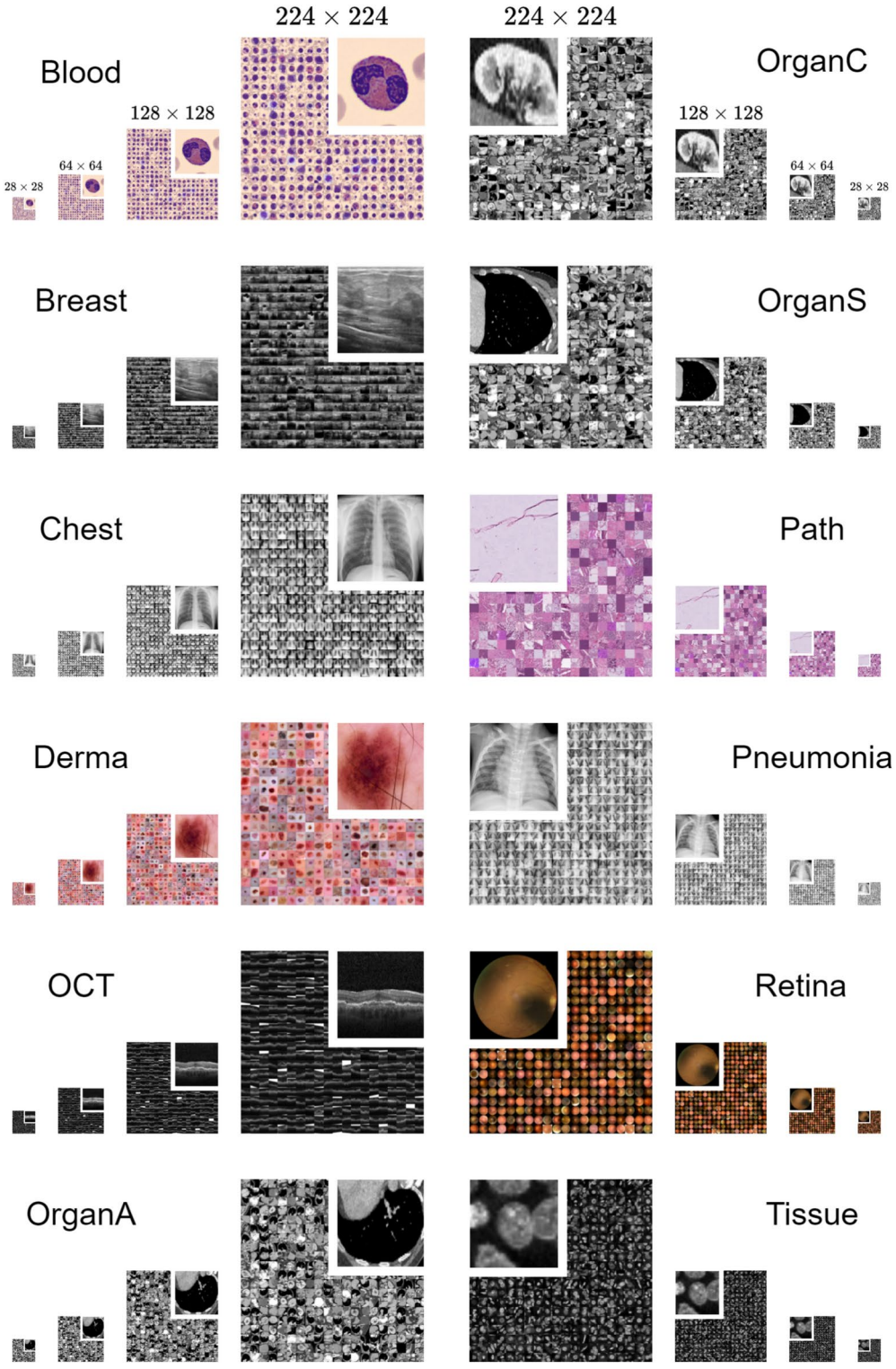


Fig. 1. Side-by-side comparison of the 12 2D datasets included in MedMNIST+, showcasing diverse primary data modalities and classification tasks across four image resolutions.

gap between expensive end-to-end training and more computationally efficient schemes such as *k*-NN or even close it entirely. On the contrary, the results for SAM illustrate that a model pretrained for a distant task (i.e. segmentation) may not readily adapt to a new task (i.e. classification) and may require complete retraining to perform effectively.

Methods	Accuracy (ACC)				Area Under the ROC Curve (AUC)			
	28×28	64×64	128×128	224×224	28×28	64×64	128×128	224×224
End-to-End								
VGG16	82.34 ± 0.88	85.33 ± 1.16	86.64±0.82	86.70±0.93	92.66 ± 0.41	94.24 ± 0.28	95.16 ± 0.27	***95.30±0.22***
AlexNet	78.92±0.81	82.94±0.78	85.04±0.74	85.74±0.64	91.14±0.43	92.72±0.33	94.29±0.30	94.90±0.23
ResNet-18	79.66±0.74	83.42±0.65	85.73±0.66	86.22±0.58	90.92±0.28	92.49±0.50	93.91±0.27	94.51±0.24
DenseNet-121	80.32±0.93	84.62±0.80	***87.13±0.56***	87.11 ± 0.64	91.75±0.55	93.59±0.23	94.57±0.21	95.03±0.23
EfficientNet-B4	73.18±1.61	79.37±1.10	82.52±0.79	82.44±1.11	87.04±0.82	90.07±0.65	91.89±0.39	91.64±0.73
ViT-B/16	78.23±0.88	83.17±0.92	84.94±0.93	86.06±0.92	90.54±0.47	92.53±0.69	93.25±0.35	94.08±0.38
CLIP ViT-B/16	76.73±0.80	80.39±0.99	82.33±1.02	82.75±1.01	89.22±1.11	90.91±0.51	91.51±0.31	91.83±0.58
EVA-02 ViT-B/16	76.69±1.44	80.77±0.97	82.76±0.97	84.72±1.09	88.91±1.01	90.53±0.54	91.59±0.75	92.60±0.94
DINO ViT-B/16	78.51±1.09	82.13±1.02	84.31±1.05	84.84±1.08	91.02±0.48	91.94±0.32	92.91±0.36	93.90±0.73
SAM ViT-B/16	78.26±0.84	82.13±1.12	84.19±1.06	84.30±0.82	89.36±0.79	90.79±1.01	91.94±1.08	91.91±0.55
Linear Probing								
VGG16	71.18±0.13	75.14±0.26	78.58±0.15	79.62±0.18	87.70±0.06	89.55±0.06	91.94±0.04	92.47±0.05
AlexNet	69.63±0.35	76.11±0.25	79.08±0.17	81.02±0.18	85.91±0.13	89.06±0.11	91.78±0.07	93.18±0.05
ResNet-18	64.41±0.05	70.49±0.08	74.94±0.07	76.89±0.08	82.95±0.18	85.24±0.22	87.97±0.76	90.37±0.17
DenseNet-121	72.10±0.28	78.01±0.19	80.77±0.19	82.22±0.23	86.76±0.94	90.11±0.29	92.00±0.19	93.02±0.11
EfficientNet-B4	67.29±0.54	73.95±0.51	76.39±0.25	77.91±0.62	83.67±0.64	86.48±0.45	88.49±0.55	89.57±0.22
ViT-B/16	73.21±0.24	79.62±0.33	83.08±0.70	84.01±0.27	88.25±0.14	91.33±0.22	93.57±0.09	94.31±0.13
CLIP ViT-B/16	74.17±0.24	78.67±0.32	81.54±0.24	82.24±0.18	88.66±0.08	91.48±0.20	93.28±0.10	93.66±0.10
EVA-02 ViT-B/16	73.04±0.19	76.50±0.16	78.48±0.11	79.30±0.10	88.41±0.04	90.66±0.03	92.04±0.04	92.41±0.04
DINO ViT-B/16	**78.23±0.22**	**82.74±0.36**	**84.46±0.24**	**85.11±0.55**	**90.94±0.14**	**93.29±0.12**	**94.50±0.11**	**94.99±0.14**
SAM ViT-B/16	43.69±0.01	48.10±0.03	54.14±0.04	61.20±0.03	66.51±1.38	74.68±0.80	80.46±0.13	81.75±0.11
k-NN ($k = 11$)								
VGG16	66.07	70.65	72.26	73.78	-	-	-	-
AlexNet	67.47	72.14	74.56	76.22	-	-	-	-
ResNet-18	66.98	71.42	74.20	76.60	-	-	-	-
DenseNet-121	67.12	71.29	74.97	76.97	-	-	-	-
EfficientNet-B4	68.15	72.45	73.91	74.17	-	-	-	-
ViT-B/16	65.92	70.90	75.45	77.51	-	-	-	-
CLIP ViT-B/16	66.40	71.61	74.83	75.52	-	-	-	-
EVA-02 ViT-B/16	69.63	72.64	74.64	75.52	-	-	-	-
DINO ViT-B/16	**73.61**	**79.41**	**81.17**	**81.90**	-	-	-	-
SAM ViT-B/16	65.05	70.40	71.58	71.95	-	-	-	-

Table 2. Benchmark outcomes summarizing the average mean and standard deviation of accuracy (ACC), for a fixed operating point of 0.5, and area under the receiver operating characteristic curve (AUC) across all datasets for all training scheme-model-image resolution combinations, derived from three independent random seeds. Notably, the k-NN algorithm, devoid of a training phase, remains unaffected by the stochasticity inherent in model training, thus reporting only the total ACC value without standard deviation. Moreover, owing to its direct utilization of embeddings and labels for classification, k-NN does not furnish a reliable AUC score. The overall best result across all training schemes, models, and resolutions is highlighted with a bold italic; the best result per resolution across all training schemes and models is highlighted with underline; and the best result per training scheme and resolution is highlighted in **bold**.

Input resolution impact

We further investigate the effect of input resolution on model performance. Our objective is to determine how often model performance enhances with incremental increases in input resolution. Intuitively, higher input resolutions are expected to enable models to capture more intricate features, potentially leading to improved overall performance. However, as demonstrated in Section Benchmark evaluation, this trend reaches a saturation point around an input resolution of 128×128 pixels for our underlying setting. To validate this observation, we analyze the instances where a model's performance surpasses that of the previous, lower resolution for each training scheme individually. Specifically, we compare the mean accuracy values across three different random seeds for the same model and training scheme between two resolutions. Figure 3 depicts this analysis for transitions from 28×28 to 64×64, 64×64 to 128×128, and 128×128 to 224×224 resolutions. We

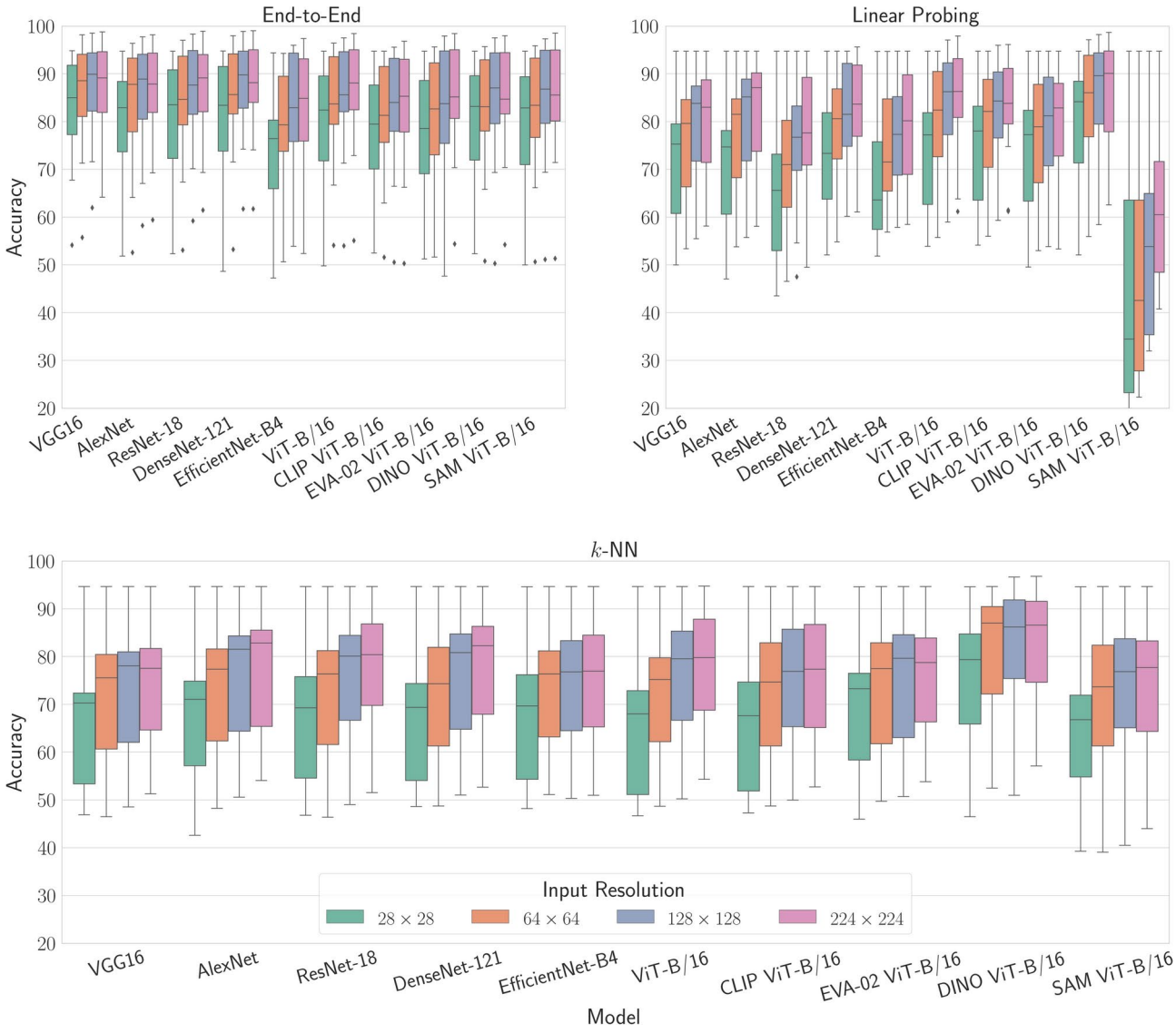


Fig. 2. Illustrating the accuracy (ACC) distributions exhibited by each model averaged across all 12 datasets, delineated by training scheme and input resolution. Each subplot within the figure illustrates the performance distributions pertaining to distinct training schemes, with color coding employed to signify the associated input resolution.

calculate the frequency of performance improvements per increase in input resolution across all 12 datasets, resulting in a maximum improvement count of 12 for each transition.

Our findings substantiate the observations from Section Benchmark evaluation to some extent. Overall, higher input resolutions lead to performance improvements across all models and training schemes. However, this improvement diminishes notably when transitioning from a 128×128 to 224×224 resolution, with superior performance observed only for a limited number of dataset instances. This trend is particularly pronounced for end-to-end trained convolutional models, whereas ViT-based models demonstrate less sensitivity to input resolution variations. This discrepancy could be attributed to the specific design of the ViT architecture, which is tailored for 224×224 pixel images, unlike convolutional models. Additionally, we observe that linear probing benefits the most from higher resolution images, with slight differences noted for the k-NN approach, potentially due to the pretraining with images of the same size. These results underscore that while input resolution impacts model performance, the effect is less significant than initially anticipated, with slight variations depending on the architecture used. This supports the utility of lower input resolutions at least during the prototyping phase.

Model ranking

We further assess how frequently a model's performance ranks among the top-5 performers concerning accuracy (ACC). Figure 4 visually portrays this as heatmaps, illustrating the total count of top-5 rank appearances for each model across all datasets, training schemes and image resolutions. Sub-figure (a) consolidates the overall ranking across all training schemes and resolutions, while sub-figure (b) presents the ranking for each training scheme

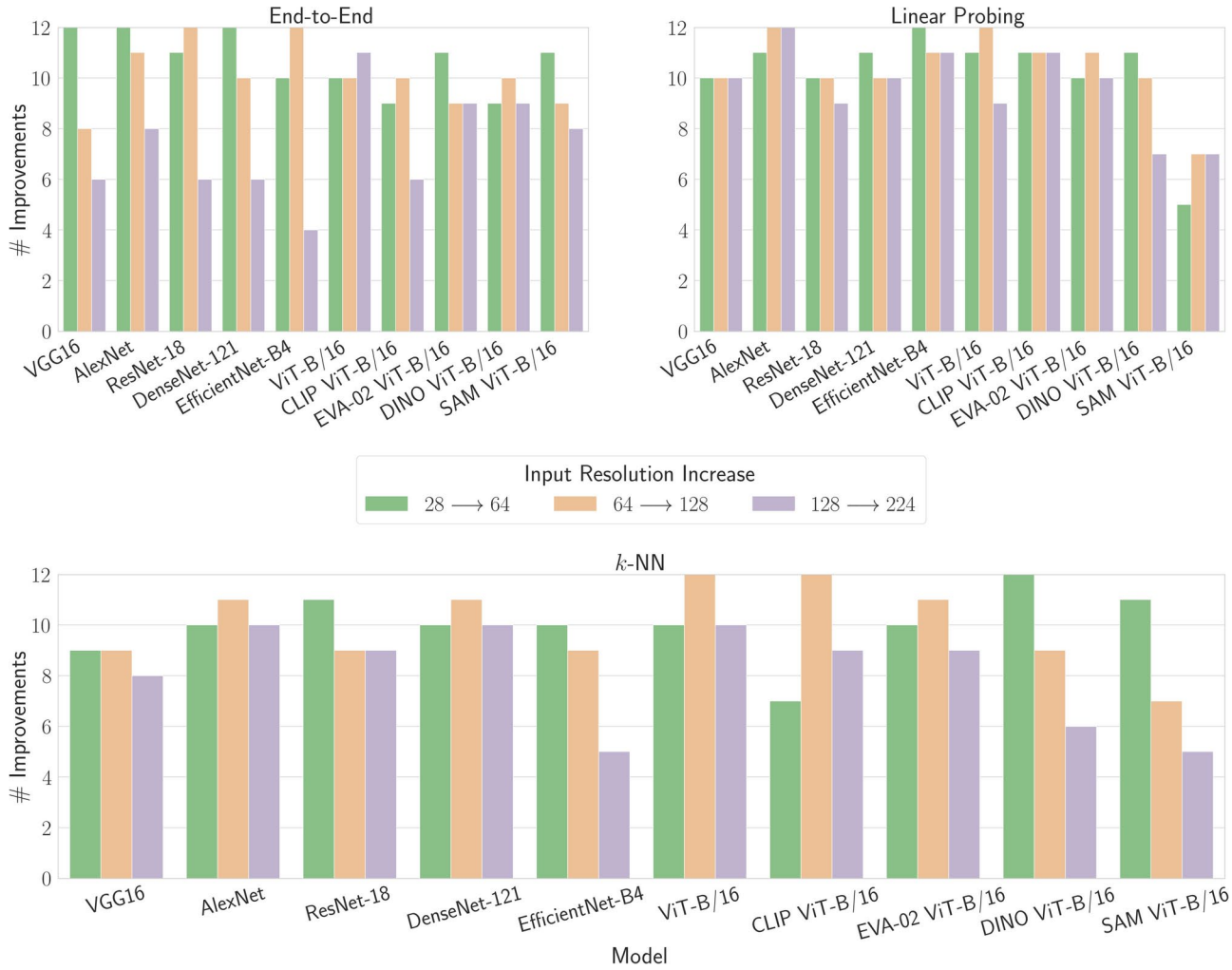


Fig. 3. Analysis of model performance (ACC) improvement with increasing input resolution across all 12 datasets. The figure illustrates the frequency of performance enhancements as input resolutions progress from 28×28 to 64×64, 64×64 to 128×128, and 128×128 to 224×224, encompassing all models and training schemes. Each bar signifies for how many datasets the model performance, in terms of the mean accuracy across the three random seeds, is superior at the next higher resolution compared to the preceding lower one, with a maximum of 12 improvements per transition.

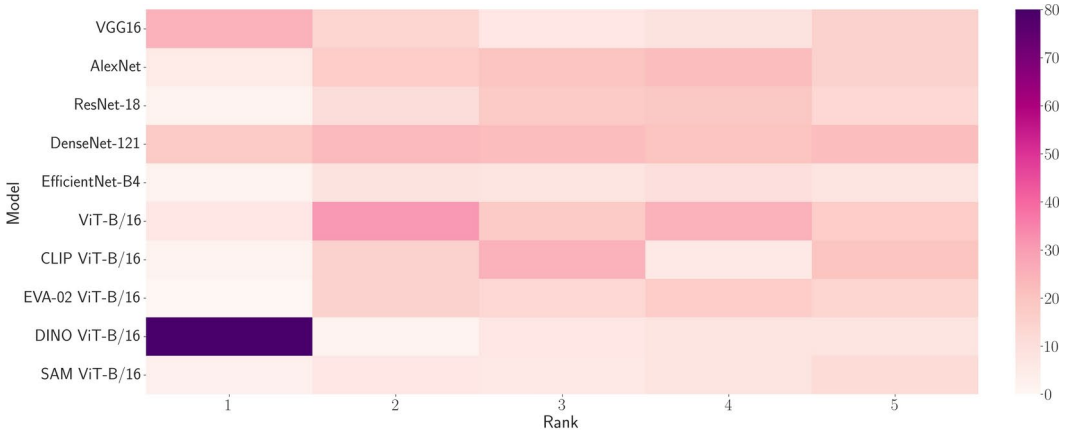
separately. At last, sub-figure (c) provides the ranking broken down on both training schemes and resolutions collectively.

Our observations unveil that convolutional models consistently outperform ViT-based models concerning ACC for end-to-end training, regardless of their pretraining strategy. Notably, VGG16 and DenseNet-121 emerge as the top performers in this aspect. The performance of the DenseNet-121 backbone is particularly intriguing, given its relatively low number of parameters and activations compared to almost all other models. This finding challenges the prevailing assumption that a more complex architecture invariably outperforms a simpler, smaller one given sufficient training samples.

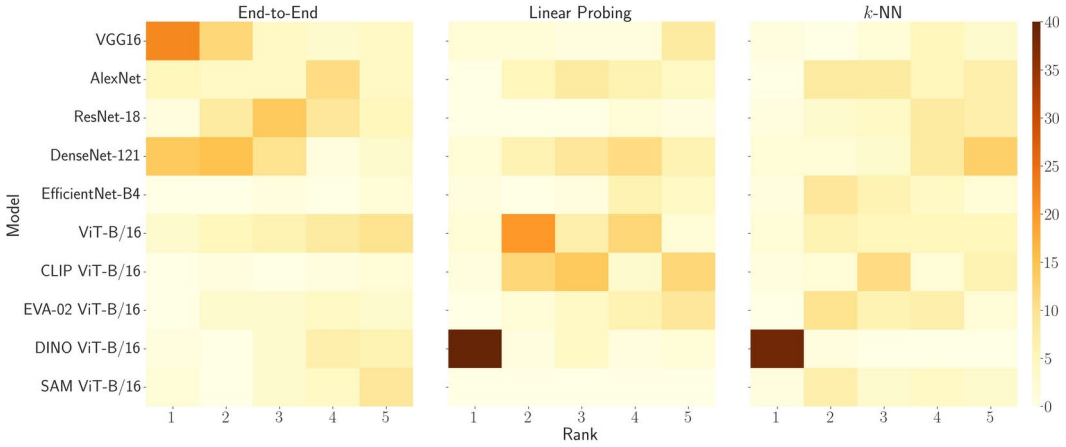
In contrast, ViT-based architectures, especially those pretrained with DINO, exhibit superior performance for linear probing and the k-NN approach compared to Convolutional Neural Networks. This is likely due to their enhanced representational capacity within the feature space. Moreover, sub-figure (c) illustrates the consistency of these observations across different input resolutions, with minimal variations observed. This underscores the significance of exhaustive pretraining for linear probing and the k-NN approach, highlighting ViTs' suitability as foundation models compared to their convolutional counterparts. Additionally, it emphasizes that the complexity of a model architecture does not necessarily align directly with its suitability for this purpose.

Quantitative evaluation

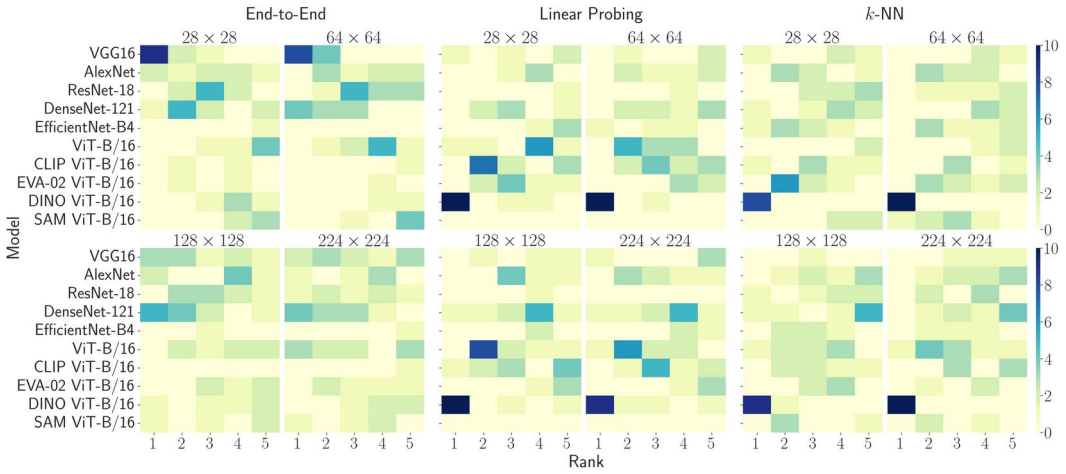
In order to substantiate the qualitative findings presented earlier, we conduct a quantitative evaluation using non-parametric statistical tests. The aim is to discern the potential impact of input resolution (specifically, 28×28 , 64×64, 128×128, and 224×224) and training scheme (comprising end-to-end training, linear probing, and k-NN integration) on the model performance, measured in terms of accuracy. Additionally, we aim to assess



(a) Frequency of top-5 performance placement per model counted for each dataset, training scheme and resolution.



(b) Frequency of top-5 performance placement per model and training scheme counted for each dataset and resolution.



(c) Frequency of top-5 performance placement per model, training scheme and resolution counted for each dataset.

Fig. 4. Ranking analysis showcasing the frequency of model placements among the top-5 performers in terms of accuracy (ACC) across all training schemes and resolutions (**a**), for each training scheme separately (**b**), and for both training schemes and resolutions, respectively (**c**) across all datasets.

whether there are notable variations in performance across datasets depending on the model architecture. To this end, Friedman tests, with a significance level set at $p = 0.05$, are initially employed to ascertain if statistically significant differences exist among the experimental conditions. Post-hoc two-tailed Wilcoxon signed-rank tests are then conducted to identify specific group differences, with a Bonferroni adjustment applied to mitigate the risk of type I errors resulting from multiple comparisons. Despite the limited number of observations ($n = 10$

for model-related evaluations and $n = 12$ for dataset-related analyses), we decided to assume an approximately normally distributed sampling distribution of the sample mean, following the Central Limit Theorem, in order to allow the computation of Z-values for the Wilcoxon tests. Finally, we asses the effect size of each test based on Cohen's interpretation guidelines[48] for Kendall's Coefficient of Concordance W, and Pearson's correlation coefficient r for the Friedman and Wilcoxon tests, respectively.

First, we want to investigate whether the input resolution significantly influences model performance. For this, we average the accuracy results per model and resolution across all datasets and training schemes. The Friedman test reveals a statistically significant difference in accuracy between the different image resolutions $\left(\chi^2(dof = 3) = 30.0,\ p = 0.000001\right)$ with a perfect agreement among the rankings of model performance across the range of input resolutions ($W = 1.0$) for 3 degrees of freedom ($dof = \text{number of input resolutions} - 1 = 3$). Median interquartile range (IQR) perceived accuracy for resolution 28×28, 64×64, 128×128, and 224×224 are 72.44 (70.76 to 73.17), 76.97 (75.60 to 77.69), 79.36 (78.38 to 80.61), and 80.10 (79.86 to 81.82), respectively. Post-hoc analysis with two-tailed Wilcoxon signed-rank tests and a Bonferroni correction, adjusted significance level of $p < 0.008$, shows significant differences between all resolution pairs ($Z = -2.803,\ p = 0.002,\ r = 0.886$). Notably, the results affirmed the trend observed in previous sections, indicating that higher resolutions generally lead to improved accuracy, albeit with diminishing returns at higher resolution levels (i.e. the transition from 128×128 to 224×224).

Next, we explore potential disparities in model performance based on the employed training schemes (end-to-end, linear probing, and k-NN). For this, we average the accuracy results per model and training scheme across all datasets and input resolutions. The Friedman test (with $dof = \text{number of training schemes} - 1 = 2$) reveals overall significant differences suggesting a strong effect of the used training scheme on the model performance, $\left(\chi^2(\text{dof} = 2) = 14.6,\ p = 0.0007,\ W = 0.73\right)$. Furthermore, post-hoc Wilcoxon signed-rank tests with Bonferroni correction, adjusted significance level of $p < 0.0167$, confirm these findings between end-to-end training and linear probing ($Z = -2.701,\ p = 0.0039,\ r = 0.854$) as well as end-to-end training and k-NN ($Z = -2.803,\ p = 0.002,\ r = 0.886$). These results, in conjunction with median IQR perceived accuracy values for end-to-end training of 82.77 (81.48 to 83.61), linear probing of 76.65 (74.45 to 78.94) and k-NN of 72.11 (72.37 to 72.60), respectively, affirm the superiority of end-to-end training in achieving the highest overall performance from Section Benchmark evaluation. Interestingly, no significant differences are observed between linear probing and k-NN integration ($Z = -1.682,\ p = 0.1055$), despite a large effect size ($r = 0.532$). This suggests comparable efficacy of these training schemes despite variations in accuracy, thereby highlighting the potential of training-free strategies such as k-NN.

Finally, we assess performance variations across datasets based on the model architecture. Through averaging accuracy results per dataset and model across all training schemes and input resolutions, we discern distinct performance trends. The IQR perceived accuracy for all models is visualized in Table 3 and the existence of statistical differences in dataset performance depending on the model architecture is illustrated in Table 4. Notably, DenseNet-121 and DINO ViT-B/16 emerged as the top-performing models, aligning with our previous findings of Section Model ranking. Except of these, ViT-based models generally exhibit superior performance compared to convolutional models, which may seem contradictory to earlier findings. However, considering that we average the performance results across all training schemes, ViT models benefit from their higher performance for linear probing and k-NN integration compared to convolutional models, which only demonstrate higher performance in end-to-end training. Thus, these findings are indeed in line with our earlier observations. Furthermore, the results for SAM confirm its inferior performance compared to all other models as depicted in Section Benchmark evaluation, thus contributing to a comprehensive understanding of model capabilities across various training contexts and tasks. Despite its similar performance for the end-to-end training, its task-foreign pretraining for segmentation rather than classification seems to limit its capabilities for training-less (linear probing) or training-free (k-NN) approaches.

	Percentile		
Model	**25th**	**50th (Median)**	**75th**
VGG16	69.78	82.54	85.45
AlexNet	69.87	**83.88**	86.00
ResNet-18	68.82	80.53	84.24
DenseNet-121	**72.04**	82.63	**87.73**
EfficientNet-B4	67.96	76.33	84.27
ViT-B/16	71.87	82.53	87.57
CLIP ViT-B/16	69.77	80.34	86.18
EVA-02 ViT-B/16	69.46	80.89	86.81
DINO ViT-B/16	**75.99**	**86.08**	**91.17**
SAM ViT-B/16	57.83	69.64	75.01

Table 3. Percentile statistics for each model performance in terms of averaged accuracy (ACC) across all training schemes and input resolutions across all 12 datasets. The highest overall value per percentile is highlighted in underline and the highest value per architecture type (convolution vs ViT, separated by a line) is highlighted in **bold**.

Model	VGG16	AlexNet	ResNet	DenseNet	EfficientNet	ViT-B/16	CLIP	EVA-02	DINO	SAM
VGG16		✖	💎	✖	✖	✖	✖	✖	✖	💎
AlexNet	✖		✖	✖	✖	✖	✖	✖	🏆	💎
ResNet	🏆	✖		🏆	✖	🏆	✖	✖	🏆	💎
DenseNet	✖	✖	💎		💎	✖	✖	✖	✖	💎
EfficientNet	✖	✖	✖	🏆		🏆	✖	✖	🏆	✖
ViT-B/16	✖	✖	💎	✖	💎		✖	💎	🏆	💎
CLIP	✖	✖	✖	✖	✖	✖		✖	🏆	💎
EVA-02	✖	✖	✖	✖	✖	🏆	✖		🏆	💎
DINO	✖	💎	💎	✖	💎	💎	💎	💎		💎
SAM	🏆	🏆	🏆	🏆	✖	🏆	🏆	🏆	🏆	

Table 4. Illustration of pair-wise significant differences between model performance in terms of averaged accuracy across all training schemes, input resolutions, and all 12 datasets using the results of the pair-wise Wilcoxon signed-rank tests with a Bonferroni correction (adjusted significance level of $p < 0.0011$). (💎: significant difference favoring the model in the **row**, 🏆: significant difference favoring the model in the **column**, ✖: no significant difference).

Further details regarding dataset-specific performance and model differences are elaborated in the appendix throughout Supplementary Tables S1 to S12.

Discussion and conclusion

This work presents a comprehensive benchmarking analysis of convolutional and Transformer-based networks, for medical image classification across diverse datasets, training schemes, and input resolutions. Through systematic evaluation, we challenge prevailing assumptions regarding model design, training schemes, and input resolution requirements. Our experiments are designed to highlight both general dataset-average findings (see Table 2) and dataset-specific results (see Supplementary Tables S1 - S12 in the appendix), which are basically coherent. By reassessing these methodologies, our aim is to foster genuine progress in the field and provide insights to inform the development of more efficient and effective models, rather than supporting the current trend of continuous scaling.

Our findings offer valuable insights into the performance of traditional models across various scenarios. End-to-end training consistently delivers the highest overall performance, with higher resolutions generally enhancing performance up to a certain threshold. Notably, we observe diminishing returns beyond 128×128 to 224×224 pixels, suggesting the potential viability of lower resolution inputs, particularly during the prototyping phase of model development. Moreover, this implies the existence of an optimal image resolution in terms of performance (accuracy and processing speed), likely lying between these two distinct image resolutions. However, this behavior is expected to be contingent on dataset characteristics, including color space and sample size. Therefore, further investigation is needed to determine the existence and dataset specificity of this optimal image resolution, as well as its contribution to overall model performance.

Furthermore, our analysis highlights the nuanced impact of self-supervised pretraining strategies like CLIP and DINO. While they do not always improve end-to-end trained models, they demonstrate enhanced performance for linear probing and k-NN integration. The near-baseline performance of DINO-pretrained models, requiring minimal training for linear probing or none for k-NN integration, raises questions about the necessity for full end-to-end training, emphasizing the potential for pretrained models to achieve comparable performance using computationally efficient methodologies. Particularly noteworthy is the fact that CLIP was trained on pairs of images and text, potentially limiting its suitability for image-centric tasks, while DINO exclusively utilizes pairs of natural images from ImageNet, which likely limits its suitability for medical image classification. The remarkable performance of CLIP and DINO despite their domain foreignness underscores the potential of foundation models and emphasizes the need for domain-specific foundation models to further enhance performance and applicability.

Finally, our model ranking analysis underscores the performance disparities between CNNs and ViTs. Convolutional models consistently outperform ViTs in accuracy for end-to-end training, while ViTs excel in linear probing and *k*-NN approaches. This emphasizes the continued competitiveness of convolutional models compared to ViTs and underscores the significance of exhaustive pretraining for the latter, highlighting the particular suitability of ViTs for foundation models.

In addition to performance considerations, robustness to distribution shifts, model interpretability, and explainability are critical factors in clinical applications, where heterogeneous data is prevalent, and trust and transparency are paramount. While our work does not explicitly evaluate these aspects, prior research suggests that ViTs and ViT-based foundation models demonstrate greater robustness to distribution shifts compared to CNNs[49–51]. Furthermore, *k*-NN integration enhances interpretability compared to end-to-end training and linear probing by enabling direct, sample-based reasoning within the feature space[52–54]. Lastly, both CNNs and ViTs can be complemented with existing explainability techniques[55,56], improving model transparency and aiding clinical decision-making. However, future work should further investigate these aspects to ensure that models not only achieve high performance but also provide clinically meaningful and interpretable insights for practitioners.

However, it is crucial to acknowledge the limitations of our study. First, our analysis is limited to the datasets and distinct dataset splits provided by MedMNIST+[33] making it inherently susceptible to biases such as class imbalance, data inhomogeneities, and demographic underrepresentation. In addition, the dataset collection lacks images from common modalities such as MRI, SPECT, and PET, and encompasses only a limited number of anatomical regions and disease patterns. Therefore, further investigation is warranted to include these modalities and assess the applicability of our findings in these unexplored settings. Second, our evaluation across all datasets simultaneously may overlook dataset-specific nuances. Future studies should explore dataset-specific results, additional datasets with varying characteristics (i.e. sample size, noisy labels, corruptions, etc.), and different dataset splits. Furthermore, our benchmark is restricted to Convolutional Neural Networks and Vision Transformer architectures. While these models remain fundamental to medical image classification, emerging architectures, such as Graph Neural Networks (GNNs), warrant further evaluation. GNNs have shown promise, particularly in histopathology, where they effectively capture topological structures in Whole Slide Images[57]. Expanding future benchmarks to include such architectures could provide a more comprehensive evaluation of the deep learning landscape for medical image classification. Additionally, we do not analyze the interpretability and explainability of the assessed architectures, which are critical for clinical adoption. Future work should address these aspects to enhance model transparency and trustworthiness. Finally, this work does not focus on the deployment of deep learning models in clinical practice which is a substantially more complex endeavor and requires addressing a broad set of challenges, including model explainability, data privacy, regulatory approval, and human-AI interaction. Instead, our benchmark aims to accelerate model development by providing insights that facilitate more efficient and scalable approaches for eventual real-world integration. In conclusion, our work advocates for the following key takeaways and recommendations for model development:

- **Prioritize computational efficiency:** prioritize the development of computationally efficient alternatives to full end-to-end training for faster model development iterations and reduced hardware demands during deployment.
- **Utilize lower-resolution images for prototyping:** consider utilizing lower resolution images during prototyping to conserve computational resources and time.
- **Benchmark models across diverse datasets:** evaluate methods on multiple distinct benchmarks to cover real-world situations, rather than focusing solely on achieving state-of-the-art performance on a single benchmark.
- **Focus on efficiency and robustness:** focus on the development of efficient and robust methods, rather than scaling existing methods to attain state-of-the-art performance.

Method

Model selection

Our selection of model architectures encompasses a diverse array of both convolutional and Transformer-based networks. Among the chosen convolutional models are well-established architectures such as VGG16[58], AlexNet[59], ResNet-18[60], DenseNet-121[61], and EfficientNet-B4[62], all of which were pretrained on the ImageNet1k dataset[63]. In the domain of Transformers, we include the Vision Transformer (ViT)[3], renowned for its exceptional performance across various image classification tasks, pretrained on ImageNet1k as well as the CLIP[4] and DINO[5] pretrained ViT variants. Acknowledging recent advancements in this area, our selection extends to adaptations of ViT such as EVA-02[64] and the Segment Anything Model (SAM)[30]. EVA-02 represents a series of efficiently optimized plain ViTs with moderate model sizes, employing bidirectional visual representations learned from a robust CLIP encoder. Conversely, SAM, initially conceived as a foundational model for image segmentation, is intentionally designed and trained to be promptable, thus facilitating zero-shot transferability to new image distributions and tasks, including image classification. For all ViT architectures, we opted for the base backbone with a patch size of 16 (i.e. ViT-B/16). With the exception of the AlexNet model obtained from the torchvision library, all models were sourced from the "Pytorch Image Models (timm)" library at Huggingface[65]. Further details regarding the employed backbone architectures, including the number of parameters, activations, Giga Multiply-Add Operations (GMACs), feature dimension before the final classification layer, memory requirements for single precision (fp32), and inference times on both GPU and CPU, are outlined in Table 5.

Model	Number Parameters (M)	Number Activations (M)	Number GMACs #	Output Dimension #	Memory Requirements (MB)	CPU Inference (ms)	GPU Inference (ms)
VGG16	138.4	13.6	15.5	4096	1454	108.0	0.4
AlexNet	62.3	0.6	0.36	4096	766	8.6	0.3
ResNet-18	11.7	2.5	1.8	512	602	9.2	0.7
DenseNet-121	8.0	6.9	2.9	1024	690	318.0	3.9
EfficientNet-B4	19.3	17.1	1.5	1792	728	29.3	3.9
ViT-B/16	86.6	16.5	16.9	768	900	49.0	1.3
CLIP ViT-B/16	86.6	16.5	16.9	768	900	48.9	1.3
EVA-02 ViT-B/16	86.3	16.5	16.9	768	902	60.0	2.8
DINO ViT-B/16	85.8	16.5	16.9	768	900	49.1	1.2
SAM ViT-B/16	89.7	64.3	23.3	256	924	71.4	13.2

Table 5. Details of evaluated model architectures including the number of parameters (in million, M), activations (in million, M), Giga Multiply-Add Operations per Second (GMACs), feature dimension before the final classification layer, memory requirements (in megabytes, MB) for single precision (fp32), and inference times on both GPU and CPU (in milliseconds, ms). GMACs, activations, inference times, and memory requirements are reported for processing a single input image of resolution 224×224 pixels. Inference times were measured on an Intel(R) Core(TM) i9-14900K processor (CPU) and an NVIDIA RTX 6000 GPU (Ada Generation).

Training pipeline

We adopt diverse training paradigms, encompassing both end-to-end training (i.e. training the whole model), and linear probing, where we solely train the classification head, while keeping the encoder frozen. Additionally, we explore the integration of the k-nearest neighbors (k-NN) classifier into the feature space of pre-trained models. Following K. Nakata et al.[66] and S. Doerrich et al.[53], the pre-trained image encoder initially extracts feature embeddings from the training set, which are then stored in an external database along with their corresponding labels. During inference, the image encoder generates a feature embedding for a given query image. Subsequently, the top-k feature embeddings having highest similarity scores (i.e. lowest cosine distance) with the query embedding are retrieved from the training set along with their associated labels. The classification of the query image is afterward determined through a majority vote on these labels. This approach facilitates efficient classification without necessitating retraining of the classification head or the entire encoder, thereby enhancing computational efficiency, interpretability, and generalizability, with reduced dependence on hyperparameters. Given the substantial computational cost associated with training deep neural networks, particularly foundation models, the adoption of k-NN methods presents an efficient alternative to traditional training schemes.

The training regimen consisted of 100 epochs with early stopping based on the validation set. We employed the AdamW optimizer[67] with a learning rate of 0.0001, along with a cosine annealing learning rate scheduler[68] with a single cycle. Each model was trained with a batch size of 64, allowing for training on a single NVIDIA RTX™ A5000 GPU. For evaluations utilizing the k-nearest neighbors (k-NN) approach, we set k to 11, in line with Z. Zhu et al.[69], who demonstrated the suitability of $k > 10$ for detecting noisy labels. To maintain compatibility with the pretrained models while preserving the inherent properties of individual resolutions, all image resolutions were padded to 224×224 pixels using zero padding. This choice was motivated by M. Hashemi[70], who demonstrated that zero-padding has no discernible effect on classification accuracy while significantly reducing training time compared to image resizing. With zero-padding, neighboring zero input units (pixels) do not activate their corresponding convolutional unit in the subsequent layer, resulting in decreased requirements for updating synaptic weights on outgoing links and ensuring robust feature preservation during image reshaping.

Loss criterion and evaluation metrics

In line with the methodology outlined by J. Yang et al.[32], we select the choice of loss criteria to suit the specific classification tasks associated with each dataset. For binary (BC) and multi-class classification (MC), as well as ordinal regression (OR) tasks, we utilize the Cross-Entropy (CE) loss function applied to the logits:

$$\mathrm{CE} = -\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{\exp(z_{n,y_n})}{\sum_{c=1}^{C}\exp(z_{n,c})}\right), \tag{1}$$

where N denotes the number of samples in the current batch, C represents the total number of classes, $z_{n,c}$ signifies the logit for class c of the n-th sample, and z_{n,y_n} denotes the logit corresponding to the target class for the n-th sample. For binary classification (BC), this equation simplifies to Binary Cross-Entropy with $C = 2$.

Additionally, we treat the multi-label classification task of the Chest dataset as a multi-label binary classification (ML-BC) problem. Here, each class label c is addressed as a distinct binary classification task, aiming to predict the presence or absence of each class label c for a given sample n. To this end, we employ the Binary Cross-Entropy with Logits (BCEwithLogits) loss function across all class labels $c \in C$:

$$\text{BCEwithLogits} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\left[y_{n,c}\log\left(\sigma(z_{n,c})\right) + (1-y_{n,c})\log\left(1-\sigma(z_{n,c})\right)\right], \tag{2}$$

where N represents the number of samples in the current batch, $z_{n,c}$ indicates the logit for sample n and class label c, $y_{n,c}$ denotes the binary label for sample n and label c, representing the presence (1) or absence (0) of class c, and $\sigma(\cdot)$ represents the sigmoid function applied to the logit $z_{n,c}$.

Furthermore, for the purpose of simplicity and standardization, consistent with J. Yang et al.[32], we employ the same evaluation metrics including accuracy (ACC) for a fixed operating point of 0.5 and the area under the receiver operating characteristic curve (AUC) to assess the model's ability to differentiate between classes. Our choice of AUC over metrics such as sensitivity, specificity, and false positive rates is driven by its suitability for comparing different architectural choices and training paradigms. Unlike sensitivity and specificity, which require setting a fixed operating point, a process that is application-dependent and not trivial, AUC provides a holistic measure of a model's discriminatory power across all operating points. Additionally, we include accuracy due to its prevalence in the field and ease of interpretation.

To quantify the level of class imbalance within each dataset, we report Shannon's Equitability for each data split. Shannon's Equitability (EH) is a normalized measure of class distribution uniformity, derived from the Shannon Diversity Index[71]. It is computed as:

$$EH = \frac{H}{H_{\max}} = \frac{-\sum_{i=1}^{S} p_i \ln p_i}{\ln S}, \tag{3}$$

where H is the Shannon Diversity Index, S represents the total number of classes, p_i denotes the proportion of samples belonging to class i and $\ln$ stands for the natural logarithm. The maximum possible diversity, $H_{\max} = \ln S$, occurs when all classes are equally represented. Shannon's Equitability ranges from 0 (total imbalance) to 1 (perfect balance), allowing for a standardized comparison of class distribution across datasets and splits.

Data availability

The used datasets are licensed under Creative Commons Attribution 4.0 International (CC BY 4.0), except DermaMNIST under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) and publicly available here: Yang, J. et al. [MedMNIST+] 18x Standardized Datasets for 2D and 3D Biomedical Image Classification with Multiple Size Options: 28 (MNIST-Like), 64, 128, and 224. https://zenodo.org/records/10519652 (2024).

Code availability

The source code is available at https://github.com/sdoerrich97/rethinking-model-prototyping-MedMNISTPlus.

Received: 24 May 2024; Accepted: 25 February 2025
Published online: 05 March 2025

References

1. Wang, Y., Liu, L. & Wang, C. Trends in using deep learning algorithms in biomedical prediction systems. *Frontiers in Neuroscience* **17** (2023).
2. Vaswani, A. *et al.* Attention is all you need. In *Neural Information Processing Systems* (2017).
3. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
4. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021).
5. Caron, M. *et al.* Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 9630–9640 (2021).
6. Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision (2024). 2304.07193.
7. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* **1**, e271–e297 (2019).
8. Sendak, M. P. *et al.* A path for translation of machine learning products into healthcare delivery. *EMJ Innovations* (2020).
9. Kocak, B., Baessler, B., Cuocolo, R., Mercaldo, N. & dos Santos, D. P. Trends and statistics of artificial intelligence and radiomics research in radiology, nuclear medicine, and medical imaging: bibliometric analysis. *European Radiology* **33**, 7542–7555. https://doi.org/10.1007/S00330-023-09772-0/FIGURES/6 (2023).
10. Stacke, K., Eilertsen, G., Unger, J. & Lundstrom, C. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics* **25**, 325–336 (2021).
11. Lafarge, M. W., Pluim, J. P., Eppenhof, K. A., Moeskops, P. & Veta, M. Domain-adversarial neural networks to address the appearance variability of histopathology images. *Lecture Notes in Computer Science* **10553 LNCS**, 83–91 (2017).

12. Oksuz, I. et al. Deep learning-based detection and correction of cardiac mr motion artefacts during reconstruction for high-quality segmentation. *IEEE Transactions on Medical Imaging* **39**, 4001–4010 (2020).
13. Khan, A. et al. Impact of scanner variability on lymph node segmentation in computational pathology. *Journal of Pathology Informatics* **13**, 100127 (2022).
14. Li, B. *et al.* Learning invariant representations and risks for semi-supervised domain adaptation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1104–1113 (2021).
15. Li, D., Yang, Y., Song, Y. Z. & Hospedales, T. M. Learning to generalize: Meta-learning for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence* **32**, 3490–3497 (2018).
16. Eche, T., Schwartz, L. H., Mokrane, F. Z. & Dercle, L. Toward generalizability in the deployment of artificial intelligence in radiology: Role of computation stress testing to overcome underspecification. *Radiology: Artificial Intelligence* **3** (2021).
17. Raji, I. D., Denton, E., Bender, E. M., Hanna, A. & Paullada, A. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021).
18. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine 2022 5:1* **5**, 1–8 (2022).
19. Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. *Electronic Markets* **31**, 685–695 (2021).
20. Koch, B., Denton, E., Hanna, A. & Foster, J. G. Reduced, reused and recycled: The life of a dataset in machine learning research. In Vanschoren, J. & Yeung, S. (eds.) *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1 (Curran, 2021).
21. Crawford, K. & Paglen, T. Excavating ai: the politics of images in machine learning training sets. *AI and Society* **36**, 1105–1116 (2021).
22. Birhane, A. & Prabhu, V. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1536–1546 (IEEE Computer Society, Los Alamitos, CA, USA, 2021).
23. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing bias in big data and ai for health care: A call for open science. *Patterns* **2**, 100347 (2021).
24. Rae, J. W. *et al.* Scaling language models: Methods, analysis & insights from training gopher. *ArXiv* **abs/2112.11446** (2021).
25. Thoppilan, R. *et al.* Lamda: Language models for dialog applications (2022). 2201.08239.
26. McKenzie, I. *et al.* The inverse scaling prize (2022).
27. Sevilla, J. *et al.* Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8, https://doi.org/10.1109/IJCNN55064.2022.9891914 (2022).
28. Goyal, A. & Bengio, Y. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **478**, https://doi.org/10.1098/rspa.2021.0068 (2022).
29. Bommasani, R. et al. *On the opportunities and risks of foundation models* **2108**, 07258 (2022).
30. Kirillov, A. et al. *Segment anything* **2304**, 02643 (2023).
31. Girdhar, R. *et al.* Imagebind: One embedding space to bind them all (2023). 2305.05665.
32. Yang, J. et al. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data* **10**, 41 (2023).
33. Yang, J. *et al.* [MedMNIST+] 18x Standardized Datasets for 2D and 3D Biomedical Image Classification with Multiple Size Options: 28 (MNIST-Like), 64, 128, and 224, https://doi.org/10.5281/zenodo.10519652 (2024).
34. Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine* **29**, 141–142 (2012).
35. Jones, R. M. *et al.* Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. *npj Digital Medicine* **3**, 144, https://doi.org/10.1038/s41746-020-00352-w (2020).
36. U.S. Food and Drug Administration. Artificial intelligence and machine learning (ai/ml)-enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (2024). Accessed: 2025-02-11.
37. Acevedo, A. et al. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief* **30**, 105474 (2020).
38. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data in Brief* **28**, 104863 (2020).
39. Wang, X. *et al.* 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)* 3462–3471 (2017).
40. 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).
41. Codella, N. C. F. *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)*, 168–172 (2018).
42. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122-1131.e9 (2018).
43. Bilic, P. et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis* **84**, 102680 (2023).
44. 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**, 1885–1898 (2019).
45. Kather, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine* **16**, e1002730 (2019).
46. Liu, R. et al. Deepdrid: Diabetic retinopathy–grading and image quality estimation challenge. *Patterns* **3**, 100512 (2022).
47. Ljosa, V., Sokolnicki, K. L. & Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nature Methods 2012 9:7* **9**, 637–637 (2012).
48. Cohen, J. Statistical power analysis. *Current Directions in Psychological Science* **1**, 98–101 (1992).
49. Zhang, C. *et al.* Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7277–7286 (2022).
50. Guo, L. L. et al. Ehr foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports* **13**, 3767. https://doi.org/10.1038/s41598-023-30820-8 (2023).
51. Zhou, A., Wang, J., Wang, Y.-X. & Wang, H. Distilling out-of-distribution robustness from vision-language foundation models. In Oh, A. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 36, 32938–32957 (Curran Associates, Inc., 2023).
52. Klein, K., De Candido, O. & Utschick, W. Interpretable classifiers based on time-series motifs for lane change prediction. *IEEE Transactions on Intelligent Vehicles* **8**, 3954–3961. https://doi.org/10.1109/TIV.2023.3276650 (2023).
53. Doerrich, S., Archut, T., Salvo, F. D. & Ledig, C. Integrating knn with foundation models for adaptable and privacy-aware image classification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5, https://doi.org/10.1109/ISBI56570.2024.10635560 (2024).
54. Ye, X., Leake, D., Wang, Y., Zhao, Z. & Crandall, D. Towards network implementation of cbr: Case study of a neural network k-nn algorithm. In Recio-Garcia, J. A., Orozco-del Castillo, M. G. & Bridge, D. (eds.) *Case-Based Reasoning Research and Development*, 354–370 (Springer Nature Switzerland, 2024).

55. Haar, L. V., Elvira, T. & Ochoa, O. An analysis of explainability methods for convolutional neural networks. *Engineering Applications of Artificial Intelligence* **117**, 105606. https://doi.org/10.1016/j.engappai.2022.105606 (2023).
56. Choi, H., Jin, S. & Han, K. Icev 2: Interpretability, comprehensiveness, and explainability in vision transformer. *International Journal of Computer Vision* https://doi.org/10.1007/s11263-024-02290-6 (2024).
57. Brussee, S., Buzzanca, G., Schrader, A. M. & Kers, J. Graph neural networks in histopathology: Emerging trends and future directions. *Medical Image Analysis* **101**, 103444 (2025).
58. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR* **abs/1409.1556** (2014).
59. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L. & Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates, Inc., 2012).
60. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2015).
61. Huang, G., Liu, Z. & Weinberger, K. Q. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269 (2016).
62. Tan, M. & Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv* **abs/1905.11946** (2019).
63. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211–252 (2014).
64. Fang, Y. *et al.* Eva-02: A visual representation for neon genesis (2023). 2303.11331.
65. Wightman, R. Pytorch image models. https://github.com/huggingface/pytorch-image-models (2019).
66. Nakata, K. *et al.* Revisiting a knn-based image classification system with high-capacity storage. In *Computer Vision – ECCV 2022*, 457–474 (Springer Nature Switzerland, 2022).
67. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2017).
68. Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with restarts. *ArXiv* **abs/1608.03983** (2016).
69. Zhu, Z., Dong, Z. & Liu, Y. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning* (2021).
70. Hashemi, M. Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data* **6**, 1–13 (2019).
71. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (1948).

Acknowledgements
Funded through the Hightech Agenda Bayern (HTA) of the Free State of Bavaria, Germany.

Author contributions
S.D., F.D.S., J.B., and C.L. worked on the methods and did the study design. S.D., F.D.S., and J.B. implemented the algorithms. J.B. analyzed the individual data sets in detail. S.D. and F.D.S. conceived and conducted the experiments. S.D. analyzed the results and created all images. C.L. supervised the study. S.D., F.D.S., and C.L. wrote the manuscript. All authors reviewed, corrected, and approved the paper.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval
This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open-access data.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-92156-9.

Correspondence and requests for materials should be addressed to S.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.