Rethinking Model Prototyping through the MedMNIST+ Dataset Collection

Sebastian Doerrich[1] **Francesco Di Salvo**[1] **Julius Brockmann**[1,2] **Christian Ledig**[1]
[1] xAILab, University of Bamberg, Germany
[2] Ludwig Maximilian University of Munich, Germany
sebastian.doerrich@uni-bamberg.de

Abstract

The integration of deep learning based systems in clinical practice is often impeded by challenges rooted in limited and heterogeneous medical datasets. In addition, prioritization of marginal performance improvements on a few, narrowly scoped benchmarks over clinical applicability has slowed down meaningful algorithmic progress. This trend often results in excessive fine-tuning of existing methods to achieve state-of-the-art performance on selected datasets rather than fostering clinically relevant innovations. In response, this work presents a comprehensive benchmark for the MedMNIST+ database to diversify the evaluation landscape and conduct a thorough analysis of common convolutional neural networks (CNNs) and Transformer-based architectures, for medical image classification. Our evaluation encompasses various medical datasets, training methodologies, and input resolutions, aiming to reassess the strengths and limitations of widely used model variants. Our findings suggest that computationally efficient training schemes and modern foundation models hold promise in bridging the gap between expensive end-to-end training and more resource-refined approaches. Additionally, contrary to prevailing assumptions, we observe that higher resolutions may not consistently improve performance beyond a certain threshold, advocating for the use of lower resolutions, particularly in prototyping stages, to expedite processing. Notably, our analysis reaffirms the competitiveness of convolutional models compared to ViT-based architectures emphasizing the importance of comprehending the intrinsic capabilities of different model architectures. Moreover, we hope that our standardized evaluation framework will help enhance transparency, reproducibility, and comparability on the MedMNIST+ dataset collection as well as future research within the field. Code is available at github/sdoerrich97.

1 Introduction

In recent years, significant strides in deep learning (DL) have reshaped various domains, from image classification to natural language processing [Wang et al., 2023]. This progress was driven by the development of increasingly sophisticated models, exemplified by architectures like the Transformer [Vaswani et al., 2017] for text or Vision Transformer (ViT) [Dosovitskiy et al., 2021] for images. Moreover, advanced training methodologies, including self-supervised contrastive methods such as CLIP [Radford et al., 2021] for image and text pairs, and DINO [Caron et al., 2021, Oquab et al., 2024] 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 [Liu et al., 2019] and software applications are beginning to receive clinical certifications [Sendak et al., 2020]. Despite this progress and the exponential growth of

Absolute link to source code: github.com/sdoerrich97/rethinking-model-prototyping-MedMNISTPlus

DL-related publications across various medical fields in the past few years [Kocak et al., 2023], the adoption of DL algorithms in daily clinical practice has been comparatively slow [Stacke et al., 2021].

One major obstacle is the scarcity of appropriate datasets, often characterized by limited sample sizes and heterogeneous image acquisition conditions [Lafarge et al., 2017, Oksuz et al., 2020, Khan et al., 2022], 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 [Li et al., 2021] or acquiring domain-invariant features [Li et al., 2018]. However, the generalizability of these methods across diverse domains remains a significant challenge, constraining their real-world applicability Eche et al. [2021].

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 [Raji et al., 2021] rather than addressing clinically relevant needs [Varoquaux and Cheplygina, 2022]. This trend is particularly pronounced in academic research, 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 [Janiesch et al., 2021]. While benchmark datasets play a crucial role in coordinating machine learning research and facilitating standardized evaluations [Koch et al., 2021], 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 [Crawford and Paglen, 2021, Birhane and Prabhu, 2021]. The latter, in particular, limits the applicability of current DL techniques across diverse patient populations, thereby impeding their real-world deployment [Norori et al., 2021]. 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 [Raji et al., 2021].

Furthermore, the limitations of scaling alone are becoming increasingly evident, as larger models start to falter in model trustworthiness [Rae et al., 2021, Thoppilan et al., 2022] or performance on well-specified tasks [McKenzie et al., 2022]. 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 [Sevilla et al., 2022]. 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 Goyal and Bengio [2022], 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 [Bommasani et al., 2021]. 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 [Kirillov et al., 2023, Girdhar et al., 2023]. 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 [Yang et al., 2023], using the recently introduced MedMNIST+ database [1]. MedMNIST v2 offers a collection of 12 distinct biomedical 2D datasets, ranging from Chest-X-ray to Dermatology, in a MNIST-like [Deng, 2012] 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

[1] MedMNIST+ including official dataset splits: `https://zenodo.org/records/10519652`

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. 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.

2 Method

2.1 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 [Simonyan and Zisserman, 2014], AlexNet [Krizhevsky et al., 2012], ResNet-18 [He et al., 2015], DenseNet-121 [Huang et al., 2016], and EfficientNet-B4 [Tan and Le, 2019], all of which were pretrained on the ImageNet1k dataset [Russakovsky et al., 2014]. In the domain of Transformers, we include the Vision Transformer (ViT) [Dosovitskiy et al., 2021], renowned for its exceptional performance across various image classification tasks, pretrained on ImageNet1k as well as the CLIP [Radford et al., 2021] and DINO [Caron et al., 2021] pretrained ViT variants. Acknowledging recent advancements in this area, our selection extends to adaptations of ViT such as EVA-02 [Fang et al., 2023] and the Segment Anything Model (SAM) [Kirillov et al., 2023]. 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 [Wightman, 2019] at Huggingface.[2][3][4] Further details regarding the employed backbone architectures, including parameter counts, the number of activations, Giga Multiply-Add Operations per Second (GMACs), and feature dimension before the last classification layer, are outlined in Table 1.

[2] timm at Hugginface: `https://huggingface.co/timm`

[3] Model identifier convolution: `vgg16.tv_in1k`, `resnet18.a1_in1k`, `densenet121.ra_in1k`, `efficientnet_b4.ra2_in1k`

[4] Model identifier transformer: `vit_base_patch16_224.augreg2_in21k_ft_in1k`, `vit_base_patch16_clip_224.laion2b_ft_in12k_in1k`, `eva02_base_patch16_clip_224.merged2b_s8b_b131k`, `vit_base_patch16_224.dino`, `samvit_base_patch16.sa1b`

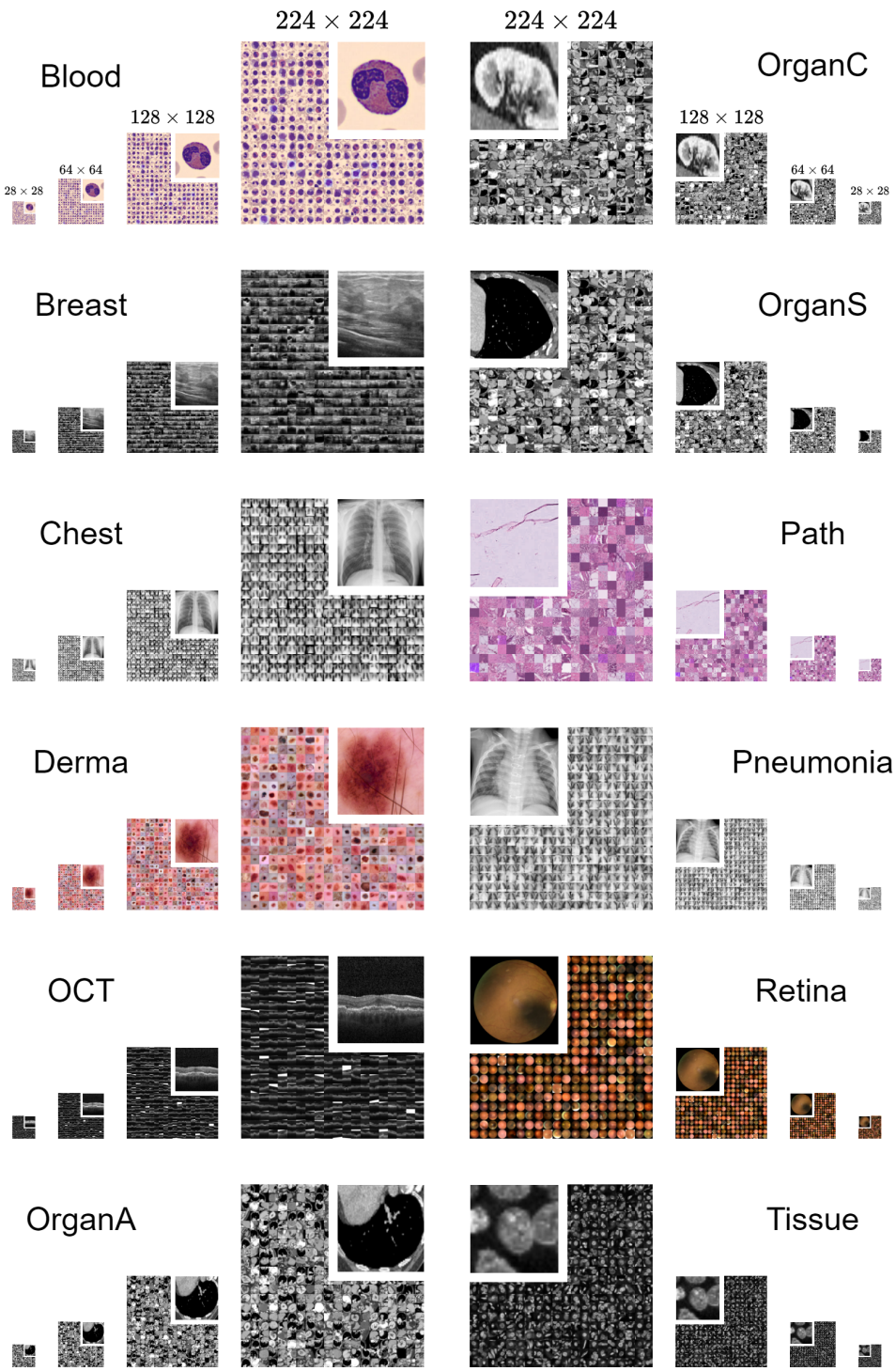


Figure 1: Side-by-side comparison of the 12 2D datasets included in MedMNIST+, showcasing diverse primary data modalities and classification tasks across four image resolutions.

Table 1: Details of evaluated model architectures including parameter count, number of activations, Giga Multiply-Add Operations per Second (GMACs) and feature dimension before the last classification layer. The number of parameters and activations is given in million (M). Except for `CLIP ViT-B/16` and `EVA-02 ViT-B/16` (both CLIP pretrained), `DINO ViT-B/16` (DINO pretrained) and `SAM ViT-B/16` (pretrained on SA-1B), all models were pretrained on ImageNet1K.

Model	Params (M)	Activations (M)	GMACs	# Output Dimension
VGG16[3]	138.4	13.6	15.5	4096
AlexNet[3]	62.3	0.6	0.36	4096
ResNet-18[3]	11.7	2.5	1.8	512
DenseNet-121[3]	8	6.9	2.9	1024
EfficientNet-B4[3]	19.3	34.8	3.1	1792
ViT-B/16[4]	86.6	16.5	16.9	768
CLIP ViT-B/16[4]	86.6	16.5	16.9	768
EVA-02 ViT-B/16[4]	86.3	16.5	16.9	768
DINO ViT-B/16[4]	85.8	16.5	16.9	768
SAM ViT-B/16[4]	89.7	1343.3	486.4	256

2.2 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 Nakata et al. [2022] and Doerrich et al. [2024], 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 [Loshchilov and Hutter, 2017] with a learning rate of 0.0001, along with a cosine annealing learning rate scheduler [Loshchilov and Hutter, 2016] 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 Zhu et al. [2021], 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 Hashemi [2019], which 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.

2.3 Loss Criterion and Evaluation Metrics

In line with the methodology outlined by Yang et al. [2023], 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:

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

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 Yang et al. [2023], we employ the same evaluation metrics including accuracy (ACC) and the area under the receiver operating characteristic curve (AUC) to assess the model's ability to differentiate between classes.

3 Experiments and Results

3.1 Datasets

The dataset selection employed in this work originates from MedMNIST v2 [Yang et al., 2023], initially introduced at a resolution of 28×28 pixels and recently expanded by MedMNIST+ 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, data modality, type of the classification task (plus number of classes) as well as the publicly available data splits, provided by MedMNIST[1]and forming a one-to-one correspondence with our benchmark, are described in Table 2. In addition, Figure 1 illustrates the datasets side-by-side for all four image resolutions.

3.2 Benchmark Evaluation

We report the performance assessment of each model selected from our diverse pool, following the methodology outlined in Section 2.2. 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 3. In addition, detailed performance evaluations for each dataset individually can be found in the appendix, spanning Tables 6 to 17. 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.

Table 2: Dataset details including data source, data modality, type of the classification task (and number of classes) as well as data splits. (ML: Multi-Label, MC: Multi-Class, BC: Binary-Class, OR: Ordinary Regression).

Dataset	Source	Modality	Task (# Classes)	# Train / Val / Test
Blood	Acevedo et al. [2020]	Blood Cell Microscope	MC (8)	11,959 / 1,712 / 3,421
Breast	Al-Dhabyani et al. [2020]	Breast Ultrasound	BC (2)	546 / 78 / 156
Chest	Wang et al. [2017]	Chest X-Ray	ML-BC (2)	78,468 / 11,219 / 22,433
Derma	Tschandl et al. [2018] Codella et al. [2019]	Dermatoscope	MC (7)	7,007 / 1,003 / 2,005
OCT	Kermany et al. [2018]	Retinal OCT	MC (4)	97,477 / 10,832 / 1,000
OrganA	Bilic et al. [2023] Xu et al. [2019]	Abdominal CT	MC (11)	34,561 / 6,491 / 17,778
OrganC	Bilic et al. [2023] Xu et al. [2019]	Abdominal CT	MC (11)	12,975 / 2,392 / 8,216
OrganS	Bilic et al. [2023] Xu et al. [2019]	Abdominal CT	MC (11)	13,932 / 2,452 / 8,827
Path	Kather et al. [2019]	Colon Pathology	MC (11)	89,996 / 10,004 / 7,180
Pneumonia	Kermany et al. [2018]	Chest X-Ray	BC (2)	4,708 / 524 / 624
Retina	Liu et al. [2022]	Fundus Camera	OR (5)	1,080 / 120 / 400
Tissue	Ljosa et al. [2012]	Kidney Cortex Microscope	MC (8)	165,466 / 23,640 / 47,280

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 plateau 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 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.

3.3 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 3.2, this trend reaches a plateau around an input resolution of 128×128 pixels for our underlying setting.

Table 3: Benchmark outcomes summarizing the average mean and standard deviation of accuracy (ACC) 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 background color; 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**.

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

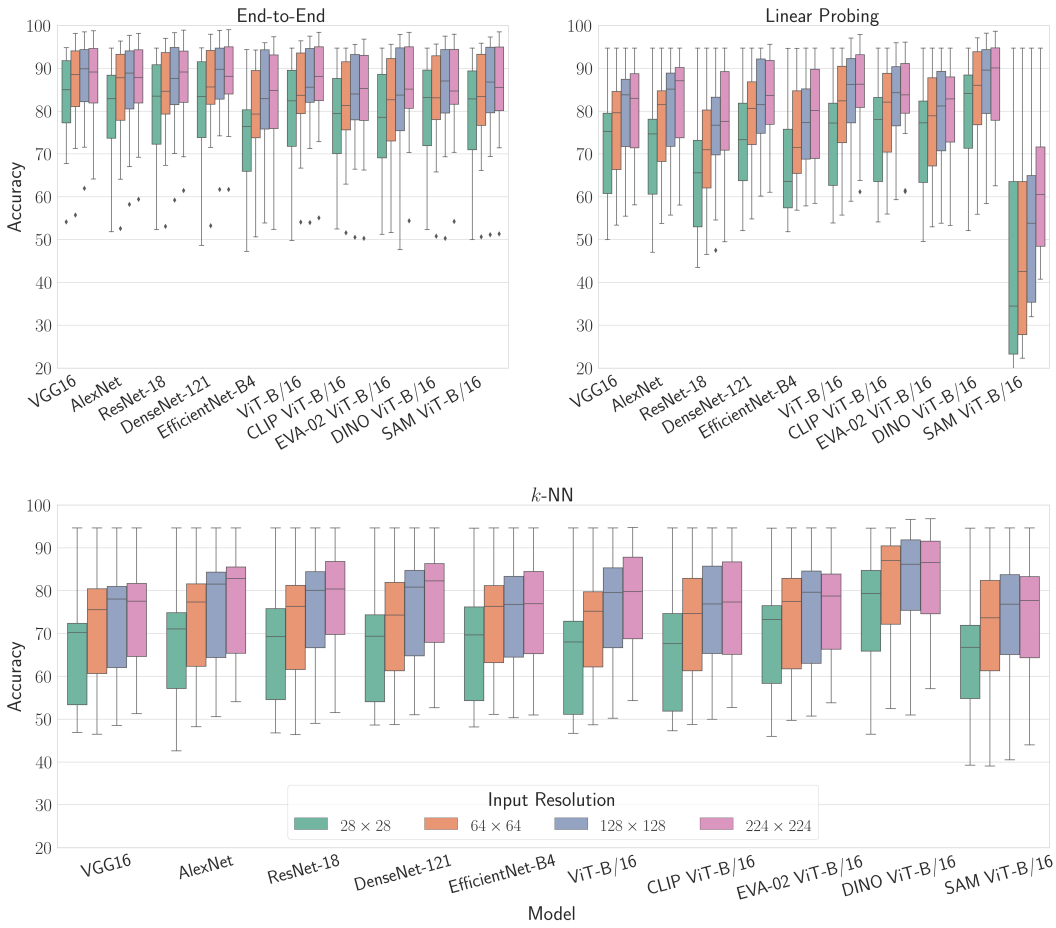


Figure 2: Illustrating the accuracy (ACC) distributions exhibited by each model averaged across all 12 datasets, delineated by training scheme and input resolution. Each subplot within the figure illustrates the performance distributions pertaining to distinct training schemes, with color coding employed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

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 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 3.2 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 exhibit 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.

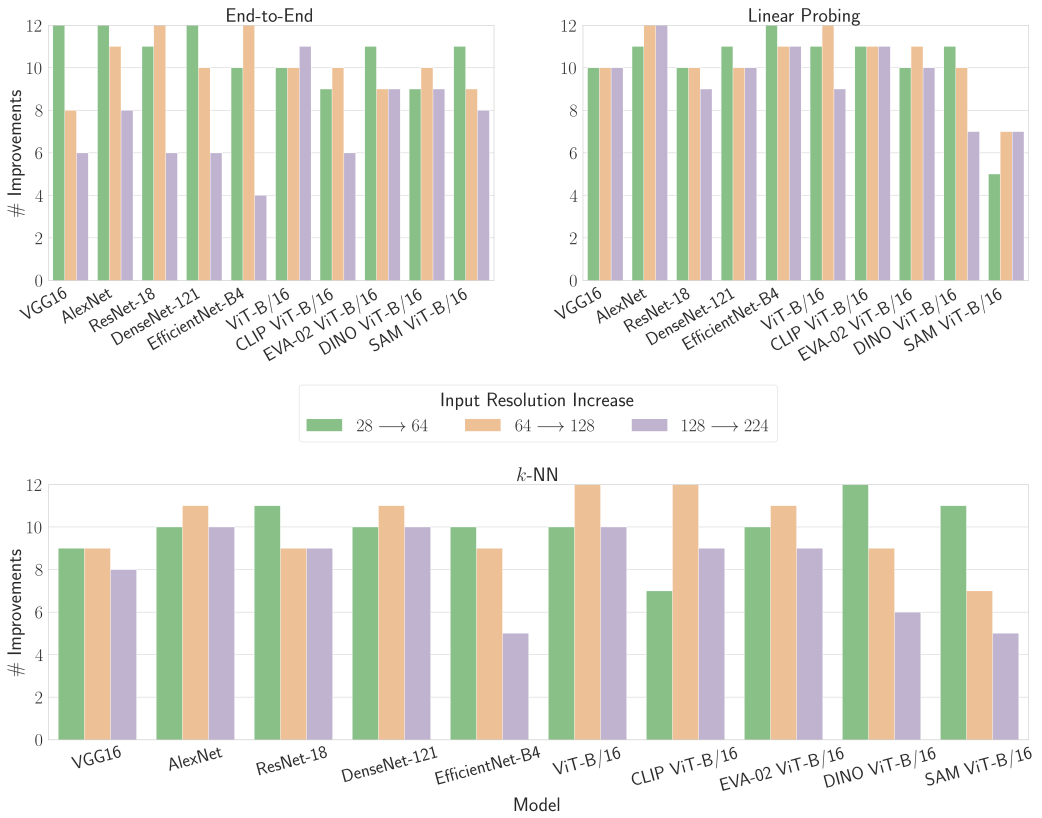


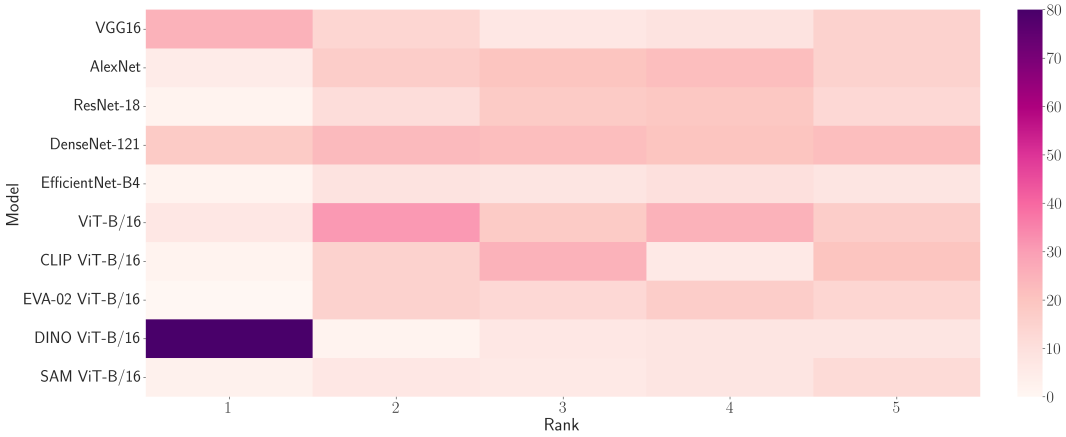
Figure 3: Analysis of model performance (ACC) improvement with increasing input resolution across all 12 datasets. The figure illustrates the frequency of performance enhancements as input resolutions progress from 28×28 to 64×64, 64×64 to 128×128, and 128×128 to 224×224, encompassing all models and 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.

3.4 Model Ranking

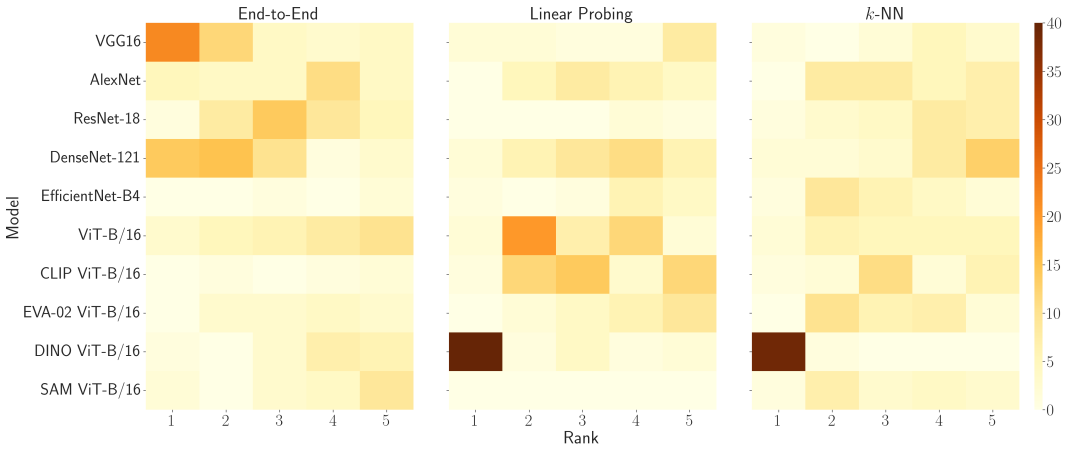
We further assess how frequently a model's performance ranks among the top-5 performers concerning accuracy (ACC). Figure 4 visually portrays this as heatmaps, illustrating the total count of top-5 rank appearances for each model across all datasets, training schemes and image resolutions. Sub-figure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

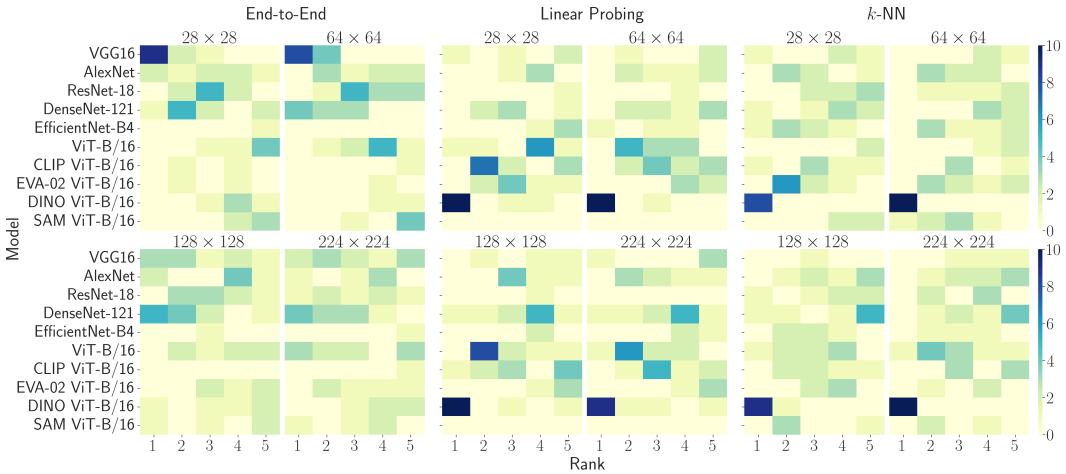
In contrast, ViT-based architectures, especially those pretrained with DINO, exhibit superior performance for linear probing and the k-NN approach compared to convolutional neural networks. This is likely due to their enhanced representational capacity within the feature space. Moreover, sub-figure (c) 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.



(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.

Figure 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.

3.5 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 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 [Cohen, 1992] 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 = 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 = 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 3.2. 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 4 and the existence of statistical differences in dataset performance depending on the model architecture is illustrated in Table 5. Notably, DenseNet-121 and DINO ViT-B/16 emerged as the top-performing models, aligning with our previous findings of Section 3.4. 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,

Table 4: 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	Percentile		
	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

the results for SAM confirm its inferior performance compared to all other models as depicted in Section 3.2, 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.

Further details regarding dataset-specific performance and model differences are elaborated in the appendix throughout Tables 6 to 17.

Table 5: 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).

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	🏆	🏆	🏆	🏆	✖	🏆	🏆	🏆	🏆	

4 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 3) and dataset-specific results (see Tables 6 - 17 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.

However, it is crucial to acknowledge the limitations of our study. Our analysis is limited to the datasets provided by MedMNIST v2 and MedMNIST+ [Yang et al., 2023], and our results are dependent on their curation and distinct dataset splits. Furthermore, 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. Additionally, 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. In conclusion, our work advocates for the following key takeaways and recommendations for model development:

- Prioritize the development of computationally efficient alternatives to end-to-end training for quicker model development iterations and reduced hardware strain during deployment.
- Consider utilizing lower resolution images during prototyping to conserve computational resources and time.
- 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 the development of efficient and robust methods, rather than scaling existing methods to attain state-of-the-art performance.

References

Yanbu Wang, Linqing Liu, and Chao Wang. Trends in using deep learning algorithms in biomedical prediction systems. *Frontiers in Neuroscience*, 17, 2023. ISSN 1662453X.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

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. In *International Conference on Learning Representations*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. 2024.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. 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(6):e271–e297, 2019. ISSN 2589-7500.

Mark P. Sendak, Joshua D'Arcy, Sehj Kashyap, Michael Gao, Marshall Nichols, Kristin M Corey, William Ratliff, Suresh Balu, Dr Sendak, Dr Gao, Dr Balu Dr, and Nichols. A path for translation of machine learning products into healthcare delivery. *EMJ Innovations*, 2020.

Burak Kocak, Bettina Baessler, Renato Cuocolo, Nathaniel Mercaldo, and Daniel Pinto dos Santos. Trends and statistics of artificial intelligence and radiomics research in radiology, nuclear medicine, and medical imaging: bibliometric analysis. *European Radiology*, 33:7542–7555, 2023. ISSN 14321084. doi: 10.1007/S00330-023-09772-0/FIGURES/6.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundstrom. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25:325–336, 2021. ISSN 21682208.

Maxime W. Lafarge, Josien P.W. Pluim, Koen A.J. Eppenhof, Pim Moeskops, and Mitko Veta. Domain-adversarial neural networks to address the appearance variability of histopathology images. *Lecture Notes in Computer Science*, 10553 LNCS:83–91, 2017. ISSN 16113349.

Ilkay Oksuz, James R. Clough, Bram Ruijsink, Esther Puyol Anton, Aurelien Bustin, Gastao Cruz, Claudia Prieto, Andrew P. King, and Julia A. Schnabel. 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. ISSN 1558254X.

Amjad Khan, Andrew Janowczyk, Felix Müller, Annika Blank, Huu Giao Nguyen, Christian Abbet, Linda Studer, Alessandro Lugli, Heather Dawson, Jean Philippe Thiran, and Inti Zlobec. Impact of scanner variability on lymph node segmentation in computational pathology. *Journal of Pathology Informatics*, 13:100127, 2022. ISSN 2153-3539.

Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao. Learning invariant representations and risks for semi-supervised domain adaptation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1104–1113, 2021. ISSN 10636919.

Da Li, Yongxin Yang, Yi Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32:3490–3497, 2018. ISSN 2374-3468.

Thomas Eche, Lawrence H. Schwartz, Fatima Zohra Mokrane, and Laurent Dercle. Toward generalizability in the deployment of artificial intelligence in radiology: Role of computation stress testing to overcome underspecification. *Radiology: Artificial Intelligence*, 3, 2021. ISSN 26386100.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 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.

Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine 2022 5:1*, 5:1–8, 2022. ISSN 2398-6352.

Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31:685–695, 2021. ISSN 14228890.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021.

Kate Crawford and Trevor Paglen. Excavating ai: the politics of images in machine learning training sets. *AI and Society*, 36:1105–1116, 2021. ISSN 14355655.

A. Birhane and V. Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society.

Natalia Norori, Qiyang Hu, Florence Marcelle Aellen, Francesca Dalia Faraci, and Athina Tzovara. Addressing bias in big data and ai for health care: A call for open science. *Patterns*, 2:100347, 2021. ISSN 26663899.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446, 2021.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. The inverse scaling prize, 2022. URL `https://github.com/inverse-scaling/prize`.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022. doi: 10.1109/IJCNN55064.2022.9891914.

Anirudh Goyal and Y. Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478, 2022. doi: 10.1098/rspa.2021.0068.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh,

Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023.

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):41, 2023.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2016.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. 2023.

Ross Wightman. Pytorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019.

Kengo Nakata, Youyang Ng, Daisuke Miyashita, Asuka Maki, Yu-Chieh Lin, and Jun Deguchi. Revisiting a knn-based image classification system with high-capacity storage. In *Computer Vision – ECCV 2022*, pages 457–474. Springer Nature Switzerland, 2022. ISBN 978-3-031-19836-6.

Sebastian Doerrich, Tobias Archut, Francesco Di Salvo, and Christian Ledig. Integrating knn with foundation models for adaptable and privacy-aware image classification, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with restarts. *ArXiv*, abs/1608.03983, 2016.

Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*, 2021.

Mahdi Hashemi. Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data*, 6:1–13, 2019. ISSN 21961115.

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.

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

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 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)*, pages 3462–3471, 2017.

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 2018 5:1*, 5:1–9, 2018. ISSN 2052-4463.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). 2019.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalena Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172:1122–1131.e9, 2018. ISSN 0092-8674.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, Jacob Sosna, Ivan Ezhov, Anjany Sekuboyina, Fernando Navarro, Florian Kofler, Johannes C. Paetzold, Suprosanna Shit, Xiaobin Hu, Jana Lipková, Markus Rempfler, Marie Piraud, Jan Kirschke, Benedikt Wiestler, Zhiheng Zhang, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Michela Antonelli, Woong Bae, Míriam Bellver, Lei Bi, Hao Chen, Grzegorz Chlebus, Erik B. Dam, Qi Dou, Chi-Wing Fu, Bogdan Georgescu, Xavier Giró i Nieto, Felix Gruen, Xu Han, Pheng-Ann Heng, Jürgen Hesser, Jan Hendrik Moltz, Christian Igel, Fabian Isensee, Paul Jäger, Fucang Jia, Krishna Chaitanya Kaluva, Mahendra Khened, Ildoo Kim, Jae-Hun Kim, Sungwoong Kim, Simon Kohl, Tomasz Konopczynski, Avinash Kori, Ganapathy Krishnamurthi, Fan Li, Hongchao Li, Junbo Li, Xiaomeng Li, John Lowengrub, Jun Ma, Klaus Maier-Hein, Kevis-Kokitsi Maninis, Hans Meine, Dorit Merhof, Akshay Pai, Mathias Perslev, Jens Petersen, Jordi Pont-Tuset, Jin Qi, Xiaojuan Qi, Oliver Rippel, Karsten Roth, Ignacio Sarasua, Andrea Schenk, Zengming Shen, Jordi Torres, Christian Wachinger, Chunliang Wang, Leon Weninger, Jianrong Wu, Daguang Xu, Xiaoping Yang, Simon Chun-Ho Yu, Yading Yuan, Miao Yue, Liping Zhang, Jorge Cardoso, Spyridon Bakas, Rickmer Braren, Volker Heinemann, Christopher Pal, An Tang, Samuel Kadoury, Luc Soler, Bram van Ginneken, Hayit Greenspan, Leo Joskowicz, and Bjoern Menze. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. ISSN 1361-8415.

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

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.

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.

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

Jacob Cohen. Statistical power analysis. *Current Directions in Psychological Science*, 1(3):98–101, 1992.

Appendix

Table 6: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the BloodMNIST dataset across 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 background color; 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**.

Methods	BloodMNIST							
	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	**94.85**±0.47	**98.20**±0.16	98.52±0.25	98.77±0.31	**99.66**±0.06	**99.91**±0.00	**99.93**±0.01	**99.95**±0.01
AlexNet	90.39±0.46	96.41±0.50	97.74±0.29	98.18±0.39	99.12±0.04	99.83±0.00	99.89±0.01	99.92±0.00
ResNet-18	91.93±0.37	97.06±0.11	98.34±0.26	98.94±0.08	99.24±0.04	99.82±0.01	99.92±0.01	99.92±0.01
DenseNet-121	93.87±0.40	97.99±0.28	**98.90**±0.09	**99.02**±0.14	99.58±0.03	**99.91**±0.02	99.92±0.01	**99.95**±0.01
EfficientNet-B4	78.06±1.34	90.61±0.39	96.00±0.19	97.40±0.27	96.62±0.26	99.26±0.05	99.78±0.03	99.90±0.00
ViT-B/16	90.59±0.56	96.42±0.26	97.54±0.21	98.43±0.09	99.21±0.08	99.80±0.00	99.91±0.01	99.93±0.01
CLIP ViT-B/16	89.92±0.25	93.56±0.90	95.60±0.42	96.82±0.04	99.08±0.03	99.56±0.04	99.78±0.03	99.87±0.01
EVA-02 ViT-B/16	91.24±0.63	95.65±0.26	97.94±0.59	98.42±0.12	99.15±0.10	99.75±0.03	99.92±0.02	99.93±0.01
DINO ViT-B/16	89.87±0.47	95.69±0.12	97.54±0.54	97.98±0.46	99.13±0.09	99.78±0.01	99.89±0.03	99.93±0.01
SAM ViT-B/16	91.19±0.54	95.89±0.12	97.33±0.32	98.55±0.26	99.16±0.07	99.72±0.01	99.86±0.04	99.92±0.02
LINEAR PROBING								
VGG16	74.95±0.02	84.53±0.04	88.86±0.09	93.77±0.09	95.66±0.00	98.24±0.00	99.10±0.00	99.64±0.00
AlexNet	66.67±0.25	81.97±0.22	89.61±0.12	91.51±0.02	93.84±0.01	97.69±0.05	99.16±0.02	99.39±0.00
ResNet-18	63.40±0.13	67.15±0.06	80.58±0.10	91.10±0.10	90.58±0.05	93.67±0.01	97.15±0.00	99.25±0.00
DenseNet-121	71.34±0.12	84.03±0.07	93.36±0.02	95.65±0.06	94.08±0.03	97.99±0.01	99.55±0.00	99.81±0.00
EfficientNet-B4	59.46±0.72	69.54±0.24	84.10±0.17	90.19±0.10	88.94±0.20	93.97±0.06	97.91±0.03	99.11±0.04
ViT-B/16	80.67±0.07	93.21±0.12	97.07±0.01	97.95±0.04	97.13±0.01	99.50±0.00	99.88±0.00	99.92±0.00
CLIP ViT-B/16	83.05±0.20	92.21±0.04	96.00±0.02	96.13±0.07	97.80±0.01	99.38±0.01	99.77±0.00	99.86±0.00
EVA-02 ViT-B/16	82.35±0.01	89.89±0.00	93.10±0.00	94.09±0.01	97.58±0.00	99.03±0.00	99.53±0.00	99.66±0.00
DINO ViT-B/16	**88.79**±0.04	**97.15**±0.09	**98.25**±0.10	**98.70**±0.01	**98.81**±0.00	**99.86**±0.00	**99.92**±0.00	**99.94**±0.00
SAM ViT-B/16	20.00±0.04	24.61±0.11	34.73±0.17	51.70±0.10	76.78±0.16	81.19±0.18	87.54±0.03	88.98±0.03
k-NN ($k = 11$)								
VGG16	71.76	75.97	79.22	81.38	-	-	-	-
AlexNet	68.34	82.23	87.55	89.94	-	-	-	-
ResNet-18	75.04	76.18	80.27	89.04	-	-	-	-
DenseNet-121	72.29	80.94	84.10	88.54	-	-	-	-
EfficientNet-B4	73.63	78.49	81.00	84.48	-	-	-	-
ViT-B/16	70.21	78.95	92.17	94.80	-	-	-	-
CLIP ViT-B/16	68.66	82.87	91.38	93.19	-	-	-	-
EVA-02 ViT-B/16	77.70	84.07	89.94	92.43	-	-	-	-
DINO ViT-B/16	**85.24**	**93.04**	**96.67**	**96.81**	-	-	-	-
SAM ViT-B/16	62.50	64.40	76.15	78.22	-	-	-	-

Table 7: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the BreastMNIST dataset across 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 background color; 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**.

BreastMNIST								
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	85.26±1.05	**87.82±2.09**	89.32±0.80	88.03±0.80	87.48±0.96	90.21±0.48	92.40±0.68	91.11±0.44
AlexNet	**86.54±1.89**	87.61±1.21	**90.60±1.09**	**88.46±0.52**	**89.08±2.41**	89.67±1.38	**93.97±1.36**	**93.86±0.48**
ResNet-18	83.97±1.81	83.33±0.52	85.47±1.51	87.18±0.52	87.61±1.16	87.14±1.77	87.78±0.63	89.94±0.57
DenseNet-121	83.33±2.91	85.90±0.91	85.68±0.30	87.39±0.80	86.18±3.25	**90.56±0.32**	86.63±0.55	89.67±0.57
EfficientNet-B4	76.50±2.47	74.57±2.88	76.50±2.47	74.57±1.32	75.62±1.69	74.09±4.58	76.73±2.92	70.84±5.18
ViT-B/16	82.05±0.52	81.62±2.58	82.48±0.80	83.76±1.09	84.83±0.96	84.49±4.67	83.63±1.49	86.18±0.26
CLIP ViT-B/16	77.56±3.43	76.71±0.60	79.06±0.80	80.13±1.81	75.19±10.6	77.58±2.42	78.38±0.73	77.53±2.64
EVA-02 ViT-B/16	74.79±5.50	73.08±0.00	72.44±1.38	82.91±3.86	74.88±6.72	73.85±1.78	78.36±3.35	83.08±5.07
DINO ViT-B/16	84.83±3.79	81.20±0.80	79.70±2.36	84.40±1.68	88.39±2.64	81.95±1.58	83.58±0.79	86.31±2.70
SAM ViT-B/16	82.05±0.52	77.35±6.04	82.91±2.63	81.62±2.12	80.16±4.38	79.16±5.18	82.07±7.69	78.51±1.13
LINEAR PROBING								
VGG16	78.63±0.30	77.78±0.60	84.62±0.00	80.98±0.60	83.10±0.24	81.86±0.22	89.00±0.08	85.06±0.17
AlexNet	77.56±1.05	81.62±0.60	84.40±0.30	86.11±0.80	77.39±0.43	82.64±0.38	90.38±0.35	**93.47±0.11**
ResNet-18	73.08±0.00	73.08±0.00	73.08±0.00	73.08±0.00	64.77±1.42	54.92±1.76	61.72±8.75	68.33±1.75
DenseNet-121	75.00±1.81	78.42±0.80	78.21±0.91	80.77±0.52	64.96±10.86	73.84±2.94	78.43±1.87	79.82±0.95
EfficientNet-B4	62.18±4.99	63.25±5.16	61.11±2.18	59.62±6.80	54.69±5.15	52.17±0.82	53.73±6.10	52.49±2.50
ViT-B/16	77.14±0.30	80.34±0.60	85.04±1.51	84.40±1.32	78.22±1.36	81.53±1.96	90.23±0.19	91.15±0.71
CLIP ViT-B/16	79.27±0.30	81.20±1.32	85.26±0.91	83.55±0.30	75.48±0.31	84.40±1.65	89.73±0.47	86.52±0.73
EVA-02 ViT-B/16	76.07±0.30	78.63±0.30	80.98±0.30	82.05±0.00	76.68±0.04	80.95±0.11	86.86±0.07	85.46±0.03
DINO ViT-B/16	**83.33±0.52**	**85.68±0.60**	**88.89±0.80**	**88.68±1.09**	**86.77±0.44**	**91.70±0.27**	**93.25±0.89**	93.43±0.82
SAM ViT-B/16	73.08±0.00	73.08±0.00	73.08±0.00	73.08±0.00	48.26±2.40	61.65±3.48	70.64±0.64	73.75±0.40
k-NN ($k = 11$)								
VGG16	74.36	80.77	80.77	79.49	-	-	-	-
AlexNet	81.41	81.41	82.69	82.69	-	-	-	-
ResNet-18	78.21	81.41	**85.90**	80.77	-	-	-	-
DenseNet-121	78.85	75.64	83.33	84.62	-	-	-	-
EfficientNet-B4	**83.33**	81.41	83.97	86.54	-	-	-	-
ViT-B/16	75.64	79.49	83.97	81.41	-	-	-	-
CLIP ViT-B/16	78.21	78.21	80.13	80.13	-	-	-	-
EVA-02 ViT-B/16	75.64	82.69	82.69	81.41	-	-	-	-
DINO ViT-B/16	78.21	**86.54**	85.26	**87.18**	-	-	-	-
SAM ViT-B/16	74.36	82.69	77.56	78.85	-	-	-	-

Table 8: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the ChestMNIST dataset across 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 background color; 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**.

Methods	ChestMNIST							
	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	**94.79±0.02**	**94.82±0.01**	94.79±0.06	94.83±0.01	**76.16±0.31**	78.59±0.11	80.07±0.26	80.92±0.11
AlexNet	94.75±0.01	94.77±0.00	**94.80±0.02**	94.79±0.01	72.31±0.20	74.75±0.29	77.21±0.32	78.87±0.22
ResNet-18	94.75±0.01	94.77±0.00	94.79±0.02	94.79±0.01	73.82±0.12	75.59±0.24	78.24±0.32	78.96±0.30
DenseNet-121	94.77±0.02	94.79±0.02	**94.80±0.03**	**94.84±0.03**	75.93±0.09	**78.70±0.38**	**81.21±0.10**	**82.22±0.14**
EfficientNet-B4	94.37±0.32	94.31±0.27	94.76±0.02	94.76±0.03	67.44±1.87	71.63±0.64	77.65±0.35	78.25±0.31
ViT-B/16	94.74±0.00	94.74±0.03	94.78±0.02	94.79±0.05	73.74±0.05	76.21±0.09	78.05±0.18	80.08±0.14
CLIP ViT-B/16	94.74±0.01	94.75±0.00	94.76±0.01	94.74±0.06	72.44±0.35	73.20±0.42	74.18±0.23	77.77±0.42
EVA-02 ViT-B/16	94.73±0.01	94.75±0.01	94.76±0.01	94.76±0.01	72.44±0.33	74.13±0.28	75.23±0.68	76.52±0.40
DINO ViT-B/16	94.74±0.01	94.73±0.03	94.76±0.01	94.78±0.02	73.23±0.21	75.13±0.27	76.81±0.26	78.99±0.22
SAM ViT-B/16	94.73±0.01	94.74±0.01	94.73±0.01	94.77±0.00	72.68±0.19	72.27±0.72	74.13±0.31	75.44±0.23
LINEAR PROBING								
VGG16	**94.75±0.00**	94.75±0.00	94.75±0.00	**94.76±0.00**	69.04±0.02	70.45±0.03	73.69±0.04	73.91±0.06
AlexNet	94.74±0.00	94.74±0.00	94.75±0.00	**94.76±0.00**	61.61±0.36	66.60±0.31	71.49±0.18	74.37±0.04
ResNet-18	94.74±0.00	94.74±0.00	94.74±0.00	94.74±0.00	61.51±0.07	62.56±0.36	66.63±0.18	75.12±0.01
DenseNet-121	94.74±0.00	94.74±0.00	94.73±0.00	94.75±0.00	63.14±0.07	68.16±0.13	71.66±0.04	76.26±0.02
EfficientNet-B4	94.68±0.02	94.69±0.03	94.72±0.01	94.74±0.00	55.93±2.16	58.77±4.36	68.20±0.30	74.89±0.01
ViT-B/16	94.74±0.00	94.73±0.00	94.74±0.00	94.73±0.00	68.04±0.03	71.37±0.11	74.67±0.18	76.09±0.07
CLIP ViT-B/16	**94.75±0.00**	94.72±0.00	94.74±0.00	94.75±0.00	70.65±0.03	71.22±0.01	74.41±0.15	76.29±0.15
EVA-02 ViT-B/16	94.74±0.00	94.74±0.00	94.73±0.00	94.74±0.00	68.87±0.01	70.89±0.01	70.65±0.22	72.11±0.22
DINO ViT-B/16	**94.75±0.00**	**94.76±0.00**	**94.76±0.00**	**94.76±0.01**	**71.69±0.20**	**74.41±0.03**	**77.90±0.06**	**78.88±0.05**
SAM ViT-B/16	94.74±0.00	94.74±0.00	94.74±0.00	94.74±0.00	60.42±0.03	60.24±0.01	61.11±0.04	63.71±0.02
k-NN ($k = 11$)								
VGG16	**94.67**	94.66	94.66	94.67	-	-	-	-
AlexNet	94.66	94.68	94.67	94.66	-	-	-	-
ResNet-18	94.66	94.68	94.66	94.66	-	-	-	-
DenseNet-121	94.66	94.67	94.66	94.66	-	-	-	-
EfficientNet-B4	94.65	94.67	94.68	94.67	-	-	-	-
ViT-B/16	94.66	94.66	**94.69**	**94.68**	-	-	-	-
CLIP ViT-B/16	94.66	94.66	94.67	94.67	-	-	-	-
EVA-02 ViT-B/16	94.64	94.67	94.68	94.66	-	-	-	-
DINO ViT-B/16	94.63	94.66	94.67	94.64	-	-	-	-
SAM ViT-B/16	94.64	**94.69**	**94.69**	**94.68**	-	-	-	-

Table 9: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the DermaMNIST dataset across 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 background color; 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**.

Methods	DermaMNIST							
	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	**76.58±0.26**	**79.53±1.19**	**81.80±1.29**	81.55±1.89	**92.69±0.59**	**94.36±0.40**	95.71±0.46	95.51±0.69
AlexNet	76.13±0.35	78.69±0.18	80.42±0.75	82.04±1.19	92.23±0.21	94.34±0.06	95.13±0.25	96.10±0.09
ResNet-18	73.38±0.61	76.36±0.22	79.19±0.51	82.33±0.36	88.35±0.28	91.70±0.76	93.46±0.39	95.27±0.27
DenseNet-121	74.16±0.65	76.16±0.58	81.76±0.25	**84.74±0.51**	91.11±0.61	93.05±0.24	95.71±0.19	96.26±0.06
EfficientNet-B4	68.74±0.75	71.45±0.10	73.83±0.30	76.38±1.29	83.87±0.66	87.58±0.27	89.32±0.16	91.88±0.85
ViT-B/16	74.40±1.51	77.01±1.45	80.81±0.89	82.31±1.36	90.62±2.21	93.77±0.80	**95.89±0.30**	96.28±0.67
CLIP ViT-B/16	72.97±0.15	72.44±0.19	74.73±0.42	75.31±0.28	90.32±0.12	91.55±0.45	92.51±0.14	92.59±0.26
EVA-02 ViT-B/16	73.48±0.65	75.28±0.98	76.41±0.46	77.94±0.29	90.17±0.34	91.47±0.34	92.63±0.33	93.23±0.33
DINO ViT-B/16	74.40±0.80	76.87±0.12	79.22±1.56	81.31±1.05	91.60±0.32	93.57±0.17	95.34±0.41	**96.50±0.51**
SAM ViT-B/16	73.08±0.65	74.68±0.10	76.71±1.06	77.42±0.17	87.97±0.65	88.33±0.58	90.66±1.36	92.60±0.65
LINEAR PROBING								
VGG16	72.15±0.05	73.77±0.11	75.38±0.16	75.99±0.13	87.61±0.01	88.71±0.01	90.53±0.01	92.03±0.01
AlexNet	72.40±0.12	75.41±0.11	77.11±0.15	78.90±0.23	88.77±0.02	91.33±0.04	92.81±0.03	94.01±0.06
ResNet-18	67.83±0.00	68.94±0.10	70.57±0.14	71.19±0.12	84.09±0.05	85.41±0.08	86.44±0.01	88.38±0.02
DenseNet-121	71.70±0.06	74.26±0.15	77.39±0.12	77.06±0.18	89.80±0.03	90.94±0.08	92.15±0.05	93.18±0.04
EfficientNet-B4	69.99±0.02	72.77±0.04	72.97±0.04	73.23±0.06	84.83±0.04	88.53±0.04	89.70±0.01	90.51±0.01
ViT-B/16	72.15±0.15	77.64±0.24	80.90±0.25	82.01±0.02	89.89±0.04	93.58±0.04	95.05±0.02	95.88±0.03
CLIP ViT-B/16	74.15±0.34	77.22±0.08	80.28±0.31	81.93±0.31	90.08±0.05	93.40±0.05	94.89±0.03	95.92±0.06
EVA-02 ViT-B/16	73.47±0.12	75.54±0.12	77.17±0.10	79.29±0.02	90.50±0.05	92.76±0.04	93.88±0.05	94.69±0.03
DINO ViT-B/16	**75.78±0.17**	**79.88±0.68**	**81.65±0.66**	**84.42±0.23**	**91.87±0.10**	**95.40±0.20**	**95.95±0.17**	**96.83±0.07**
SAM ViT-B/16	66.88±0.00	66.88±0.00	66.88±0.00	66.88±0.00	66.73±0.88	70.87±0.37	72.64±0.46	69.34±0.33
k-NN ($k = 11$)								
VGG16	70.27	72.27	72.57	71.67	-	-	-	-
AlexNet	70.62	73.92	74.16	74.66	-	-	-	-
ResNet-18	70.72	71.52	71.37	73.07	-	-	-	-
DenseNet-121	69.33	70.62	72.67	73.17	-	-	-	-
EfficientNet-B4	69.48	71.62	71.87	71.72	-	-	-	-
ViT-B/16	69.43	70.87	72.62	75.21	-	-	-	-
CLIP ViT-B/16	72.12	71.17	73.52	74.46	-	-	-	-
EVA-02 ViT-B/16	73.22	73.77	74.11	75.61	-	-	-	-
DINO ViT-B/16	**73.97**	**75.91**	**76.51**	**78.35**	-	-	-	-
SAM ViT-B/16	69.73	70.42	70.22	68.38	-	-	-	-

Table 10: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the OctMNIST dataset across 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 background color; 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**.

Methods	OctMNIST							
	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	**77.50±2.86**	81.93±2.44	**90.50±0.85**	**90.30±2.62**	**95.84±1.00**	97.50±0.79	98.73±0.26	99.17±0.13
AlexNet	66.40±1.13	75.47±2.28	80.57±1.90	84.63±1.73	90.96±0.39	95.09±0.76	97.92±0.22	98.01±0.48
ResNet-18	69.07±0.83	80.70±2.01	84.03±1.97	85.10±1.12	92.10±0.36	97.53±0.26	97.77±0.58	98.81±0.28
DenseNet-121	72.87±2.17	**84.40±1.93**	89.83±3.65	86.63±0.42	94.35±0.83	**98.00±0.40**	98.60±0.61	**99.26±0.24**
EfficientNet-B4	59.93±2.79	75.87±1.55	80.73±2.00	82.57±4.67	88.65±0.74	95.61±0.08	98.09±0.12	98.93±0.24
ViT-B/16	64.00±0.33	80.23±0.58	87.47±2.21	90.13±0.39	88.59±0.47	96.51±0.22	**98.95±0.22**	99.12±0.15
CLIP ViT-B/16	61.53±0.70	77.73±1.78	83.43±3.42	86.80±3.07	87.46±0.13	95.68±0.03	98.11±0.58	98.88±0.41
EVA-02 ViT-B/16	60.63±0.98	73.00±2.79	84.33±1.32	87.43±1.11	86.60±0.94	94.24±0.79	98.13±0.63	98.93±0.18
DINO ViT-B/16	64.53±1.03	78.40±1.49	84.03±1.27	85.07±2.23	89.26±0.70	96.39±0.17	98.22±0.51	98.57±0.29
SAM ViT-B/16	64.87±2.38	80.07±0.58	87.50±1.19	87.30±0.88	89.26±0.24	96.83±0.52	98.87±0.32	99.19±0.13
LINEAR PROBING								
VGG16	50.03±0.05	58.90±0.08	70.77±0.25	67.30±0.36	84.59±0.01	89.79±0.02	94.66±0.01	95.81±0.02
AlexNet	47.07±0.05	56.57±0.19	62.50±0.24	68.30±0.29	82.24±0.02	84.84±0.06	90.71±0.05	94.31±0.09
ResNet-18	46.73±0.05	53.40±0.08	68.60±0.00	72.00±0.20	83.04±0.01	88.34±0.02	96.39±0.01	97.65±0.01
DenseNet-121	56.17±0.12	66.07±0.24	71.10±0.33	78.47±0.34	88.87±0.01	94.37±0.01	97.27±0.01	98.72±0.01
EfficientNet-B4	54.17±0.05	66.23±0.05	72.17±0.05	76.53±0.05	88.94±0.01	93.56±0.00	96.44±0.01	97.73±0.00
ViT-B/16	54.43±0.05	66.80±0.33	76.87±0.12	**83.57±0.12**	86.31±0.01	94.94±0.01	97.22±0.04	**98.96±0.03**
CLIP ViT-B/16	58.20±0.16	63.37±0.26	**77.23±0.17**	81.13±0.17	89.47±0.01	93.00±0.05	97.66±0.02	98.66±0.00
EVA-02 ViT-B/16	53.33±0.17	60.27±0.05	63.87±0.19	68.50±0.24	87.44±0.02	92.72±0.02	95.95±0.01	96.47±0.01
DINO ViT-B/16	**62.63±0.09**	**71.20±0.49**	75.47±0.21	73.73±0.99	**92.30±0.04**	**95.99±0.11**	**97.83±0.04**	98.35±0.08
SAM ViT-B/16	26.20±0.00	28.87±0.05	34.50±0.00	40.80±0.00	66.00±0.05	72.17±0.06	71.19±0.03	80.36±0.03
k-NN ($k = 11$)								
VGG16	46.90	46.50	52.30	61.30	-	-	-	-
AlexNet	42.60	49.90	50.60	54.60	-	-	-	-
ResNet-18	46.80	46.40	56.00	68.90	-	-	-	-
DenseNet-121	**49.20**	50.30	55.30	63.20	-	-	-	-
EfficientNet-B4	48.20	54.20	63.60	65.80	-	-	-	-
ViT-B/16	46.70	48.70	58.90	65.70	-	-	-	-
CLIP ViT-B/16	47.30	52.30	61.40	58.90	-	-	-	-
EVA-02 ViT-B/16	46.00	51.20	55.50	58.90	-	-	-	-
DINO ViT-B/16	46.50	**61.10**	**72.20**	**74.10**	-	-	-	-
SAM ViT-B/16	39.30	39.10	40.50	44.00	-	-	-	-

Table 11: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the OrganAMNIST dataset across 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 background color; 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**.

Methods	OrganAMNIST							
	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	**93.19±0.49**	**96.44±0.35**	94.24±1.94	94.61±0.64	99.47±0.05	99.79±0.04	99.75±0.12	99.75±0.05
AlexNet	90.87±0.21	95.23±0.38	95.34±0.33	95.93±0.29	99.37±0.03	99.78±0.06	99.79±0.03	99.85±0.04
ResNet-18	92.04±0.04	95.49±0.12	96.21±0.30	96.01±0.11	99.30±0.12	99.76±0.05	99.84±0.02	99.70±0.02
DenseNet-121	91.62±0.64	96.16±0.29	**96.51±0.14**	96.72±0.29	**99.52±0.02**	**99.84±0.02**	**99.88±0.01**	99.84±0.05
EfficientNet-B4	85.81±1.30	93.52±0.67	95.58±0.23	95.11±0.27	98.76±0.09	99.65±0.08	99.85±0.02	99.79±0.01
ViT-B/16	90.29±0.62	94.78±0.94	96.27±0.55	**96.93±0.43**	99.20±0.07	99.80±0.08	99.86±0.03	**99.91±0.02**
CLIP ViT-B/16	88.10±0.23	93.90±0.34	94.83±0.43	95.25±0.39	98.97±0.06	99.70±0.06	99.81±0.03	99.81±0.05
EVA-02 ViT-B/16	88.54±0.47	93.66±1.30	95.93±0.56	96.12±0.22	98.94±0.14	99.57±0.15	99.84±0.03	99.88±0.01
DINO ViT-B/16	89.98±0.05	94.99±0.37	95.96±0.57	96.13±0.32	99.31±0.05	99.81±0.04	99.87±0.03	99.90±0.02
SAM ViT-B/16	90.41±0.32	94.42±0.39	95.96±0.29	95.60±0.19	99.00±0.08	99.59±0.07	99.75±0.03	99.78±0.02
LINEAR PROBING								
VGG16	79.36±0.09	85.62±0.37	89.09±0.08	91.46±0.04	97.49±0.02	98.75±0.06	99.25±0.01	99.53±0.00
AlexNet	79.87±0.50	90.03±0.09	92.32±0.16	93.32±0.08	97.52±0.08	99.36±0.00	99.58±0.02	99.67±0.00
ResNet-18	70.37±0.04	84.38±0.06	89.30±0.06	90.20±0.05	94.85±0.00	98.43±0.00	99.20±0.00	99.35±0.00
DenseNet-121	81.99±0.08	90.71±0.04	92.73±0.05	93.63±0.10	98.03±0.02	99.42±0.01	99.67±0.00	99.73±0.01
EfficientNet-B4	74.23±0.02	86.98±0.03	90.13±0.15	90.79±0.04	96.15±0.00	98.95±0.01	99.35±0.01	99.38±0.01
ViT-B/16	81.45±0.43	90.21±0.15	92.14±0.06	93.06±0.36	97.49±0.08	99.42±0.02	99.59±0.00	99.65±0.03
CLIP ViT-B/16	80.36±0.06	88.20±0.11	90.19±0.12	90.96±0.08	97.62±0.01	99.19±0.02	99.42±0.01	99.45±0.01
EVA-02 ViT-B/16	81.68±0.29	87.12±0.05	88.50±0.10	89.97±0.20	97.92±0.03	98.93±0.01	99.18±0.01	99.37±0.03
DINO ViT-B/16	**89.74±0.37**	**93.91±0.16**	**94.97±0.13**	**94.96±0.02**	**99.27±0.04**	**99.74±0.01**	**99.79±0.01**	**99.78±0.01**
SAM ViT-B/16	22.54±0.07	39.10±0.25	61.65±0.08	71.18±0.05	79.07±0.05	90.11±0.02	94.50±0.01	95.29±0.01
k-NN ($k = 11$)								
VGG16	70.42	80.40	82.90	84.55	-	-	-	-
AlexNet	72.71	82.87	86.54	88.38	-	-	-	-
ResNet-18	69.68	81.25	86.08	86.69	-	-	-	-
DenseNet-121	69.48	81.93	86.73	87.28	-	-	-	-
EfficientNet-B4	69.92	81.16	83.19	82.15	-	-	-	-
ViT-B/16	66.71	80.67	83.13	83.86	-	-	-	-
CLIP ViT-B/16	66.65	79.68	81.29	82.77	-	-	-	-
EVA-02 ViT-B/16	73.33	80.74	82.71	83.59	-	-	-	-
DINO ViT-B/16	**84.59**	**90.75**	**91.25**	**90.69**	-	-	-	-
SAM ViT-B/16	70.08	82.34	83.67	83.14	-	-	-	-

Table 12: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the OrganCMNIST dataset across 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 background color; 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**.

Methods	OrganCMNIST							
	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	91.35±0.41	93.51±0.72	92.54±0.88	93.37±0.34	**99.32±0.08**	99.64±0.05	99.61±0.03	99.65±0.02
AlexNet	87.76±1.73	92.82±0.25	93.84±0.20	93.17±0.16	99.13±0.16	99.61±0.02	99.72±0.03	99.71±0.03
ResNet-18	90.48±0.42	93.25±0.27	93.94±0.31	93.20±0.12	99.14±0.03	99.60±0.06	99.68±0.02	99.61±0.07
DenseNet-121	**91.52±0.33**	**93.99±0.32**	**94.42±0.37**	93.72±0.52	99.27±0.05	**99.69±0.02**	99.65±0.05	99.67±0.02
EfficientNet-B4	84.25±1.79	89.15±2.09	90.93±0.70	90.60±0.16	98.39±0.17	99.17±0.23	99.50±0.05	99.46±0.01
ViT-B/16	89.34±1.00	93.20±0.53	93.34±0.46	**94.02±0.45**	99.13±0.08	99.71±0.02	99.72±0.05	**99.78±0.04**
CLIP ViT-B/16	87.52±0.25	90.88±0.84	91.54±1.63	92.39±0.43	98.94±0.03	99.40±0.02	99.53±0.15	99.61±0.08
EVA-02 ViT-B/16	88.77±0.56	91.86±0.59	93.93±0.34	**94.02±0.84**	98.77±0.16	99.33±0.08	99.61±0.09	99.61±0.05
DINO ViT-B/16	89.51±0.75	92.36±0.09	93.78±1.39	93.68±0.14	99.24±0.01	99.63±0.02	**99.75±0.09**	**99.78±0.02**
SAM ViT-B/16	89.04±0.54	92.94±0.05	92.12±1.13	92.71±0.51	98.63±0.17	99.31±0.09	99.33±0.13	99.39±0.10
LINEAR PROBING								
VGG16	75.62±0.28	81.49±0.21	84.30±0.36	85.11±0.11	96.67±0.05	98.03±0.03	98.58±0.05	98.81±0.01
AlexNet	78.15±0.50	87.39±0.13	88.66±0.12	89.74±0.14	97.32±0.10	98.97±0.02	99.16±0.01	99.20±0.01
ResNet-18	60.70±0.02	75.86±0.12	81.65±0.03	82.21±0.01	92.56±0.01	97.04±0.00	98.20±0.00	98.33±0.00
DenseNet-121	76.22±0.10	85.80±0.09	87.70±0.10	88.09±0.13	97.01±0.00	98.88±0.00	99.14±0.01	99.21±0.01
EfficientNet-B4	65.08±0.03	78.61±0.06	81.67±0.07	83.85±0.07	93.77±0.00	97.70±0.00	98.27±0.00	98.59±0.00
ViT-B/16	77.36±0.30	85.46±0.14	90.82±5.27	88.20±0.05	97.22±0.04	98.92±0.01	99.06±0.02	99.19±0.00
CLIP ViT-B/16	76.78±0.11	83.04±0.20	83.93±0.13	84.12±0.19	96.93±0.02	98.54±0.03	98.68±0.00	98.72±0.03
EVA-02 ViT-B/16	78.51±0.04	81.63±0.29	83.60±0.34	83.85±0.10	97.41±0.00	98.36±0.05	98.66±0.04	98.66±0.00
DINO ViT-B/16	**88.34±0.06**	**92.21±0.12**	**91.82±0.22**	**92.46±1.70**	**99.08±0.01**	**99.62±0.01**	**99.60±0.02**	**99.54±0.01**
SAM ViT-B/16	22.33±0.00	22.33±0.00	32.01±0.11	58.60±0.10	62.32±3.37	87.56±0.08	93.09±0.02	93.31±0.00
k-NN ($k = 11$)								
VGG16	67.70	75.21	77.31	75.65	-	-	-	-
AlexNet	72.47	80.22	81.86	83.00	-	-	-	-
ResNet-18	66.20	76.56	80.00	80.08	-	-	-	-
DenseNet-121	62.97	73.03	78.44	79.99	-	-	-	-
EfficientNet-B4	63.83	74.25	72.69	73.47	-	-	-	-
ViT-B/16	64.74	72.59	76.06	78.25	-	-	-	-
CLIP ViT-B/16	59.30	70.87	73.69	74.66	-	-	-	-
EVA-02 ViT-B/16	72.48	75.23	76.64	77.57	-	-	-	-
DINO ViT-B/16	**83.63**	**87.52**	**87.13**	**86.00**	-	-	-	-
SAM ViT-B/16	71.15	82.05	84.07	83.80	-	-	-	-

Table 13: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the OrganSMNIST dataset across 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 background color; 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**.

Methods	OrganSMNIST							
	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	**78.87±0.68**	81.60±0.21	82.36±0.08	82.06±0.48	**97.64±0.09**	98.06±0.24	98.11±0.06	97.89±0.18
AlexNet	76.75±0.44	80.58±0.90	81.71±0.17	81.56±0.36	97.30±0.13	97.87±0.17	98.07±0.06	98.17±0.03
ResNet-18	76.24±0.36	80.34±0.82	82.30±0.29	81.24±0.42	97.23±0.15	97.91±0.07	98.19±0.13	97.86±0.15
DenseNet-121	77.70±0.61	**83.47±0.24**	**83.18±0.19**	81.80±0.66	97.17±0.19	97.93±0.05	97.94±0.13	98.11±0.16
EfficientNet-B4	67.97±0.79	76.01±0.73	77.33±0.38	76.37±0.42	95.07±0.20	97.17±0.08	97.50±0.04	97.33±0.03
ViT-B/16	76.45±0.83	81.43±1.28	82.94±0.33	**82.50±0.60**	97.12±0.03	**98.19±0.19**	**98.50±0.08**	98.31±0.18
CLIP ViT-B/16	73.35±0.31	78.08±2.76	79.65±1.72	78.69±1.71	96.53±0.19	97.56±0.33	97.96±0.15	97.76±0.11
EVA-02 ViT-B/16	71.91±3.07	78.77±1.99	81.19±1.35	81.62±0.80	95.80±0.65	97.49±0.24	98.02±0.15	98.09±0.11
DINO ViT-B/16	76.10±0.85	79.38±2.39	82.72±0.33	81.72±0.57	97.09±0.18	98.00±0.12	98.33±0.05	**98.33±0.05**
SAM ViT-B/16	76.71±0.41	80.18±0.19	80.61±0.54	81.00±0.63	96.11±0.14	97.34±0.15	97.48±0.21	97.70±0.10
LINEAR PROBING								
VGG16	63.33±0.23	68.83±0.28	72.08±0.23	72.85±0.07	93.70±0.07	95.48±0.06	96.37±0.01	96.62±0.00
AlexNet	63.75±0.77	72.14±0.38	74.91±0.04	75.65±0.18	94.44±0.16	96.73±0.06	97.24±0.00	97.12±0.00
ResNet-18	53.65±0.06	64.95±0.11	70.19±0.04	70.12±0.06	89.55±0.00	94.31±0.01	95.78±0.01	95.84±0.00
DenseNet-121	66.28±0.03	74.96±0.06	76.10±0.12	76.56±0.19	94.86±0.00	96.93±0.02	97.32±0.02	97.41±0.02
EfficientNet-B4	58.50±0.05	70.32±0.03	71.35±0.03	72.06±0.07	91.96±0.00	95.72±0.01	96.16±0.01	96.44±0.00
ViT-B/16	65.44±0.15	74.63±0.28	77.41±0.24	77.45±0.22	94.46±0.01	97.05±0.02	97.48±0.05	97.55±0.01
CLIP ViT-B/16	65.39±0.07	72.82±0.07	74.67±0.15	74.77±0.13	94.23±0.01	96.61±0.00	96.87±0.04	96.93±0.02
EVA-02 ViT-B/16	66.42±0.13	69.52±0.05	73.08±0.09	74.24±0.09	95.01±0.02	96.21±0.02	96.54±0.02	96.91±0.01
DINO ViT-B/16	**74.25±0.36**	**78.70±0.06**	**80.79±0.24**	**79.27±0.17**	**97.03±0.02**	**97.92±0.00**	**98.09±0.00**	**97.89±0.01**
SAM ViT-B/16	23.54±0.00	23.54±0.00	35.64±0.02	49.05±0.03	55.66±5.45	56.44±3.28	89.56±0.02	89.66±0.02
k-NN ($k = 11$)								
VGG16	55.25	63.80	64.98	65.80	-	-	-	-
AlexNet	60.67	66.50	68.29	68.97	-	-	-	-
ResNet-18	56.67	65.83	70.26	70.08	-	-	-	-
DenseNet-121	55.61	64.96	67.96	69.56	-	-	-	-
EfficientNet-B4	55.84	66.22	64.81	63.92	-	-	-	-
ViT-B/16	52.02	66.04	69.30	69.81	-	-	-	-
CLIP ViT-B/16	50.91	64.35	66.65	67.27	-	-	-	-
EVA-02 ViT-B/16	60.61	65.29	65.62	68.81	-	-	-	-
DINO ViT-B/16	**70.69**	**76.64**	**77.70**	**74.82**	-	-	-	-
SAM ViT-B/16	56.72	67.13	70.62	69.77	-	-	-	-

Table 14: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the PathMNIST dataset across 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 background color; 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**.

Methods	PathMNIST							
	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	**88.93±0.53**	**93.83±1.12**	94.85±0.61	94.73±0.31	**98.57±0.14**	99.34±0.12	99.42±0.18	99.33±0.06
AlexNet	80.75±1.63	89.00±1.35	93.00±0.65	94.19±0.99	96.67±0.41	98.78±0.17	99.35±0.04	99.41±0.11
ResNet-18	85.54±0.87	93.36±0.65	95.27±0.23	93.82±1.19	98.24±0.14	99.40±0.11	99.62±0.10	99.30±0.15
DenseNet-121	85.26±0.36	92.34±0.84	94.69±0.31	95.74±0.58	98.18±0.14	**99.43±0.06**	**99.70±0.06**	**99.79±0.06**
EfficientNet-B4	76.35±3.04	88.39±1.94	94.20±0.64	92.60±0.57	96.27±0.61	98.90±0.25	99.57±0.02	99.45±0.13
ViT-B/16	82.82±0.29	91.97±0.39	94.61±0.90	95.82±0.25	97.81±0.31	99.28±0.11	99.67±0.08	99.64±0.10
CLIP ViT-B/16	81.41±0.75	87.47±1.96	92.71±0.34	92.51±1.56	97.79±0.22	98.79±0.13	99.49±0.01	99.47±0.11
EVA-02 ViT-B/16	82.33±1.71	90.54±1.68	94.88±1.02	95.97±0.85	97.69±0.16	99.17±0.19	99.69±0.10	99.75±0.10
DINO ViT-B/16	82.23±1.32	90.29±1.80	94.24±0.23	94.33±0.90	97.56±0.38	99.13±0.07	99.68±0.03	99.54±0.22
SAM ViT-B/16	84.45±0.53	91.82±1.60	**95.67±1.27**	**96.07±0.41**	98.10±0.23	99.37±0.10	99.66±0.02	99.76±0.05
LINEAR PROBING								
VGG16	80.04±0.01	84.88±0.05	87.00±0.04	87.84±0.06	96.79±0.00	98.21±0.00	98.50±0.00	98.76±0.01
AlexNet	76.95±0.04	81.48±0.06	86.86±0.14	88.21±0.14	95.33±0.00	96.87±0.02	98.59±0.01	98.79±0.03
ResNet-18	73.53±0.01	84.42±0.02	88.12±0.02	88.94±0.02	95.42±0.00	98.05±0.00	98.57±0.00	98.88±0.00
DenseNet-121	81.82±0.01	90.03±0.05	92.02±0.02	91.28±0.06	97.80±0.01	99.11±0.00	99.40±0.00	99.21±0.01
EfficientNet-B4	80.51±0.03	86.45±0.02	87.62±0.02	89.67±0.02	97.33±0.00	98.74±0.00	98.84±0.00	98.88±0.00
ViT-B/16	83.90±0.04	91.43±0.02	92.70±0.02	93.54±0.08	97.90±0.00	99.33±0.00	99.50±0.00	99.63±0.01
CLIP ViT-B/16	83.78±0.02	90.83±0.08	91.12±0.17	91.82±0.07	**98.01±0.00**	99.17±0.00	99.33±0.00	99.32±0.01
EVA-02 ViT-B/16	82.53±0.05	90.07±0.04	91.73±0.03	87.37±0.03	97.76±0.00	99.03±0.00	99.34±0.00	99.02±0.00
DINO ViT-B/16	**85.05±0.11**	**93.90±0.11**	**94.27±0.07**	**96.14±0.05**	97.93±0.02	**99.59±0.01**	**99.64±0.01**	**99.75±0.00**
SAM ViT-B/16	31.28±0.05	56.31±0.01	64.34±0.05	75.75±0.01	77.04±0.02	88.94±0.01	91.06±0.00	96.37±0.00
k-NN ($k = 11$)								
VGG16	70.32	76.82	78.84	82.19	-	-	-	-
AlexNet	71.52	74.55	81.23	84.21	-	-	-	-
ResNet-18	68.89	79.25	81.85	83.47	-	-	-	-
DenseNet-121	72.90	82.16	86.16	85.86	-	-	-	-
EfficientNet-B4	73.84	80.45	81.25	80.45	-	-	-	-
ViT-B/16	71.96	81.50	86.57	88.04	-	-	-	-
CLIP ViT-B/16	73.48	83.02	85.58	86.49	-	-	-	-
EVA-02 ViT-B/16	76.17	83.58	86.49	79.94	-	-	-	-
DINO ViT-B/16	**80.54**	**90.39**	**93.72**	**94.32**	-	-	-	-
SAM ViT-B/16	63.91	76.99	78.02	77.28	-	-	-	-

Table 15: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the PneumoniaMNIST dataset across 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 background color; 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**.

Methods	PneumoniaMNIST							
	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	84.78±2.55	**89.26±0.39**	87.13±1.93	87.39±1.00	**97.04±0.54**	**98.69±0.16**	98.55±0.44	**98.47±0.16**
AlexNet	**85.10±0.73**	88.03±1.64	87.18±1.67	87.23±0.77	95.88±0.42	96.71±0.64	97.85±0.32	98.25±0.13
ResNet-18	83.12±1.10	86.00±1.44	89.85±1.37	**91.13±0.96**	95.01±0.30	94.34±0.71	97.91±0.23	98.04±0.30
DenseNet-121	83.55±2.21	85.42±1.36	89.80±0.79	88.94±2.30	96.06±0.24	97.19±0.29	**98.75±0.23**	97.41±0.80
EfficientNet-B4	79.01±1.44	82.69±0.13	85.15±0.93	87.13±1.14	90.69±0.46	95.34±0.52	96.91±0.21	96.71±0.72
ViT-B/16	83.65±2.77	85.84±1.32	83.76±2.17	86.11±3.57	95.60±0.27	96.41±0.10	96.81±0.73	96.33±0.51
CLIP ViT-B/16	84.56±1.96	84.62±1.04	84.62±1.73	83.81±1.83	94.66±0.36	94.63±0.19	94.61±0.75	95.59±1.30
EVA-02 ViT-B/16	85.04±1.52	86.59±1.14	83.17±0.32	82.75±2.53	93.88±1.04	94.99±1.14	93.81±1.21	93.54±1.82
DINO ViT-B/16	84.13±2.06	85.04±2.16	**90.06±0.65**	84.08±2.13	95.52±0.62	96.00±0.28	97.96±0.71	96.67±0.95
SAM ViT-B/16	83.71±2.08	86.65±0.98	86.16±2.69	83.81±2.85	91.94±1.45	93.68±2.74	95.88±1.61	94.53±2.71
LINEAR PROBING								
VGG16	81.57±0.26	84.13±0.32	83.33±0.23	86.22±0.13	91.93±0.03	95.59±0.10	96.15±0.03	96.75±0.04
AlexNet	78.10±0.50	83.92±0.40	85.90±0.26	88.09±0.15	91.14±0.07	94.89±0.05	97.51±0.06	97.57±0.01
ResNet-18	74.36±0.26	78.90±0.20	80.40±0.20	83.17±0.13	87.69±0.08	94.21±0.02	94.40±0.05	96.91±0.01
DenseNet-121	81.94±0.20	82.85±0.13	84.88±0.27	86.59±0.20	92.68±0.14	96.01±0.06	96.53±0.05	97.34±0.05
EfficientNet-B4	82.64±0.08	84.19±0.08	84.40±0.15	87.07±0.08	93.66±0.03	96.29±0.02	95.41±0.02	97.21±0.01
ViT-B/16	83.12±0.20	84.51±0.72	87.50±0.35	88.30±0.26	94.28±0.02	95.93±0.11	97.33±0.03	97.63±0.07
CLIP ViT-B/16	84.88±0.20	85.04±0.20	84.67±0.20	84.99±0.27	94.48±0.19	96.54±0.05	96.64±0.11	97.17±0.07
EVA-02 ViT-B/16	83.60±0.08	79.22±0.15	81.52±0.08	83.65±0.13	94.10±0.03	94.30±0.01	95.42±0.03	96.34±0.05
DINO ViT-B/16	**86.59±0.54**	**86.43±0.65**	**90.33±0.15**	**91.56±0.27**	**97.29±0.07**	**97.56±0.10**	**98.89±0.03**	**98.92±0.08**
SAM ViT-B/16	62.50±0.00	62.50±0.00	62.50±0.00	62.50±0.00	81.25±0.50	87.16±0.25	92.52±0.05	90.13±0.14
k-NN ($k = 11$)								
VGG16	76.12	83.01	81.73	81.57	-	-	-	-
AlexNet	81.89	81.25	83.65	84.62	-	-	-	-
ResNet-18	81.09	86.06	83.97	87.34	-	-	-	-
DenseNet-121	82.05	82.37	84.29	86.06	-	-	-	-
EfficientNet-B4	85.58	83.81	84.78	84.62	-	-	-	-
ViT-B/16	83.01	77.88	84.94	87.82	-	-	-	-
CLIP ViT-B/16	85.26	83.01	86.38	87.50	-	-	-	-
EVA-02 ViT-B/16	84.29	79.81	83.97	84.94	-	-	-	-
DINO ViT-B/16	**85.74**	**87.82**	**90.54**	**89.74**	-	-	-	-
SAM ViT-B/16	80.93	83.33	84.13	86.06	-	-	-	-

Table 16: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the RetinaMNIST dataset across 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 background color; 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**.

Methods	RetinaMNIST							
	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	**54.17**±0.77	**55.75**±4.48	**62.00**±1.02	**64.17**±2.42	**75.35**±0.98	**80.63**±0.68	**85.26**±0.60	**87.82**±0.71
AlexNet	51.83±0.66	52.58±0.42	58.25±1.74	59.42±0.96	72.80±0.64	75.08±0.38	80.02±0.95	83.20±1.02
ResNet-18	52.33±2.37	53.08±1.48	59.25±0.41	61.50±1.34	70.41±0.55	74.40±1.90	80.81±0.61	83.16±0.65
DenseNet-121	48.67±0.51	53.25±2.15	61.75±0.20	61.75±1.08	71.17±1.16	74.30±0.74	81.45±0.51	82.90±0.57
EfficientNet-B4	47.25±2.70	50.67±0.77	53.92±1.36	52.42±1.66	64.12±2.59	70.83±0.30	73.61±0.65	73.83±0.76
ViT-B/16	49.83±1.53	54.08±0.96	54.00±2.27	55.08±2.05	71.72±0.89	73.58±1.65	73.71±0.81	78.58±2.32
CLIP ViT-B/16	52.50±1.06	51.58±0.85	50.58±1.01	50.33±0.62	72.80±0.93	72.69±1.72	71.43±0.62	70.84±1.46
EVA-02 ViT-B/16	51.25±1.22	51.67±0.24	47.67±3.37	54.42±1.53	71.23±0.91	71.15±1.19	69.91±2.11	74.73±2.80
DINO ViT-B/16	52.33±1.64	50.83±1.90	50.33±2.01	54.25±2.41	73.32±0.28	71.91±0.70	71.96±0.60	78.32±3.37
SAM ViT-B/16	50.00±1.95	50.67±2.71	51.17±1.05	51.33±1.45	71.25±1.60	71.64±1.66	71.90±0.96	71.72±1.39
LINEAR PROBING								
VGG16	50.58±0.24	53.42±1.01	57.33±0.24	61.08±0.42	71.84±0.31	75.50±0.17	81.06±0.18	84.86±0.27
AlexNet	51.25±0.35	54.25±0.74	56.17±0.47	58.08±0.12	70.94±0.17	74.07±0.32	78.31±0.1	81.33±0.20
ResNet-18	43.50±0.00	46.58±0.24	47.50±0.20	49.50±0.20	68.91±0.52	71.18±0.35	75.52±0.16	79.45±0.21
DenseNet-121	52.08±0.77	54.83±0.59	**60.17**±0.31	**62.67**±0.92	**73.41**±0.14	77.01±0.22	**83.46**±0.24	**85.93**±0.23
EfficientNet-B4	51.83±0.42	**57.42**±0.42	58.58±0.12	58.75±0.20	72.51±0.14	76.23±0.07	80.23±0.07	81.32±0.07
ViT-B/16	**54.25**±1.08	55.75±1.27	59.00±0.54	61.17±0.72	73.09±0.07	75.15±0.32	82.47±0.52	85.16±0.55
CLIP ViT-B/16	54.17±1.45	56.00±1.41	59.33±0.66	61.25±0.54	73.10±0.29	**77.77**±0.55	82.29±0.33	85.35±0.09
EVA-02 ViT-B/16	49.58±1.03	53.00±0.89	53.83±0.12	53.33±0.31	72.45±0.32	76.86±0.14	79.75±0.04	81.05±0.15
DINO ViT-B/16	52.08±0.12	55.92±1.36	58.42±0.31	62.58±1.90	71.99±0.63	77.18±0.69	82.17±0.13	85.57±0.57
SAM ViT-B/16	43.50±0.00	43.50±0.00	43.50±0.00	43.50±0.00	56.00±3.62	66.13±1.83	65.07±0.27	63.22±0.33
k-NN ($k = 11$)								
VGG16	47.75	51.25	53.25	55.75	-	-	-	-
AlexNet	46.75	48.25	52.75	54.75	-	-	-	-
ResNet-18	47.50	49.00	51.00	53.50	-	-	-	-
DenseNet-121	49.50	48.75	**55.00**	58.00	-	-	-	-
EfficientNet-B4	49.75	52.00	54.75	51.00	-	-	-	-
ViT-B/16	48.50	48.75	50.25	56.25	-	-	-	-
CLIP ViT-B/16	**52.25**	48.75	50.00	52.75	-	-	-	-
EVA-02 ViT-B/16	51.75	49.75	50.75	54.50	-	-	-	-
DINO ViT-B/16	48.00	**52.50**	51.00	**59.00**	-	-	-	-
SAM ViT-B/16	49.25	52.00	49.75	52.25	-	-	-	-

Table 17: Benchmark outcomes summarizing the mean and standard deviation of accuracy (ACC) and area under the receiver operating characteristic curve (AUC) for the TissueMNIST dataset across 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 background color; 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**.

TissueMNIST								
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	**67.75±0.46**	71.29±0.81	71.62±0.16	70.57±0.35	**92.67±0.12**	94.12±0.23	94.37±0.10	94.05±0.06
AlexNet	59.80±0.43	64.12±0.25	67.06±0.06	69.25±0.33	88.79±0.11	91.16±0.04	92.56±0.07	93.50±0.08
ResNet-18	63.02±0.10	67.36±0.20	70.17±0.73	69.35±0.67	90.59±0.08	92.65±0.09	93.65±0.21	93.57±0.13
DenseNet-121	66.53±0.37	**71.54±0.71**	**74.25±0.39**	**74.08±0.32**	92.47±0.04	**94.46±0.24**	**95.35±0.07**	**95.25±0.04**
EfficientNet-B4	59.92±0.55	65.16±1.63	71.35±0.24	69.31±1.50	88.97±0.50	91.59±0.71	94.15±0.06	93.35±0.53
ViT-B/16	60.55±0.56	66.72±0.71	71.29±0.40	72.89±0.70	88.97±0.18	92.62±0.31	94.32±0.17	94.84±0.16
CLIP ViT-B/16	56.63±0.53	62.97±0.67	66.47±0.29	66.25±0.28	86.44±0.27	90.55±0.35	92.32±0.25	92.20±0.11
EVA-02 ViT-B/16	57.60±1.00	64.42±0.67	70.42±0.87	70.34±0.93	87.31±0.58	91.21±0.29	93.92±0.33	93.97±0.35
DINO ViT-B/16	59.44±0.31	65.80±1.01	69.35±1.72	70.38±0.99	88.55±0.27	91.98±0.44	93.48±0.77	94.02±0.37
SAM ViT-B/16	58.86±0.20	66.19±0.70	69.42±0.60	71.47±0.41	88.07±0.32	92.25±0.33	93.70±0.23	94.40±0.13
LINEAR PROBING								
VGG16	53.19±0.00	53.53±0.02	55.50±0.08	58.12±0.13	83.95±0.00	84.03±0.00	86.36±0.03	87.89±0.04
AlexNet	49.10±0.10	53.80±0.06	55.75±0.06	59.57±0.05	80.35±0.13	84.74±0.06	86.47±0.03	88.98±0.01
ResNet-18	51.06±0.00	53.43±0.02	54.60±0.02	56.46±0.01	82.44±0.00	84.76±0.00	85.65±0.00	86.97±0.00
DenseNet-121	55.93±0.01	59.46±0.02	60.86±0.02	61.09±0.02	86.45±0.00	88.64±0.00	89.38±0.01	89.64±0.01
EfficientNet-B4	54.24±0.01	56.91±0.00	57.88±0.02	58.47±0.01	85.29±0.00	87.17±0.00	87.63±0.00	88.33±0.00
ViT-B/16	53.91±0.06	60.69±0.09	62.79±0.02	63.80±0.08	84.95±0.03	89.27±0.03	90.39±0.01	90.88±0.06
CLIP ViT-B/16	55.28±0.03	59.39±0.04	61.03±0.08	61.50±0.06	86.02±0.01	88.60±0.02	89.62±0.01	89.76±0.08
EVA-02 ViT-B/16	54.23±0.01	58.35±0.02	59.71±0.00	60.51±0.02	85.14±0.00	87.93±0.00	88.68±0.00	89.24±0.00
DINO ViT-B/16	**57.46±0.22**	**63.09±0.02**	**63.92±0.05**	**64.04±0.12**	**87.22±0.10**	**90.51±0.02**	**91.02±0.01**	**90.98±0.03**
SAM ViT-B/16	37.70±0.00	41.70±0.00	46.05±0.01	46.66±0.01	68.59±0.01	73.73±0.01	76.61±0.00	76.89±0.00
k-NN ($k = 11$)								
VGG16	47.30	47.20	48.55	51.30	-	-	-	-
AlexNet	45.97	49.96	50.79	54.11	-	-	-	-
ResNet-18	48.25	48.89	49.05	51.54	-	-	-	-
DenseNet-121	48.66	50.10	51.04	52.67	-	-	-	-
EfficientNet-B4	49.79	51.14	50.35	51.26	-	-	-	-
ViT-B/16	47.47	50.73	52.84	54.33	-	-	-	-
CLIP ViT-B/16	48.04	50.48	53.29	53.46	-	-	-	-
EVA-02 ViT-B/16	49.74	50.92	52.52	53.86	-	-	-	-
DINO ViT-B/16	**51.56**	**56.10**	**57.39**	**57.12**	-	-	-	-
SAM ViT-B/16	48.00	49.67	49.64	46.97	-	-	-	-