SEUSL

Sri Lankan Journal of Technology
Research

SLJoT
2026 Issue 01 67-81
ISSN-2773-6970

Benchmarking Pre-Trained Feature Extractors: A Comparative Study Across Deep Learning Tasks

Rafeek Sibrikhan[1] and M.M. Mohamed Mufassirin[2*]

[1]Department of Computer Science, Faculty of Science, University of Jaffna
[2]Department of Computer Science, Faculty of Applied Sciences, South Eastern University of Sri Lanka

*Corresponding Author: mufassirin@seu.ac.lk || ORCID: 0000-0002-3141-7023

Received: 20-06-2026 * Accepted: 28-06-2026 * Published Online: 30-06-2026

Abstract- The rapid growth of artificial intelligence and deep learning has revolutionized diverse domains, including computer vision where transfer learning through pre-trained models has become a fundamental technique for enhancing performance of models and reducing computational costs. Although numerous pre-trained deep learning models have been widely adopted, their effectiveness varies significantly across application domains and datasets. The lack of comprehensive comparative studies makes it challenging for researchers to identify the most suitable models for achieving optimal performance in specific tasks. This study systematically compares seven pre-trained feature extractors across three architectural families, convolutional neural networks (CNNs), Vision Transformers (ViTs), and self-supervised models to provide practical guidance on model selection for downstream deep learning tasks. These models were evaluated on five benchmark datasets. Features were extracted with frozen weights and evaluated using linear probing, k-nearest neighbor retrieval, and 5-shot classification. ConvNeXt-B achieved the highest mean linear probe accuracy (90.40%), while DINOv2-S produced the best feature geometry for retrieval tasks (87.64%). CLIP-ViT-B/32 demonstrated the strongest cross-domain transfer, leading on texture recognition and satellite imagery few-shot classification. Older CNN architectures lagged significantly behind modern models by approximately 7%. Overall, this study simultaneously evaluates multiple feature extractors across multiple visual domains and evaluation protocols.

Keywords- Feature Extraction, Transfer Learning, Convolutional Neural Networks, Vision Transformers, Self-Supervised Learning

Recommended APA Citation

Sibrikhan, R., & Mufassirin, M. M. M. (2026). Benchmarking pre-trained feature extractors: A comparative study across deep learning tasks. *Sri Lankan Journal of Technology, 7*(1), 68–81.

Introduction

The current paradigm in computer vision is to use transfer learning, where pre-trained feature extractors are used. Representations learnt from large amounts of labelled data on tasks like ImageNet (Niu et al., 2025) perform well even in the case of a small amount of labelled data. The quality of the feature extractor that transforms the raw image into a small embedding vector is the critical element in this pipeline, and directly affects the performance of the downstream components. In the last 10 years, the set of feature extractors has changed considerably. The classic approach to transfer learning was using a convolutional neural network (CNN), like ResNet (Li et al., 2025) and DenseNet (Kaya & Eris, 2023), as the backbone of the model. Later developments created architectures that improved on the accuracy-efficiency balance, such as EfficientNet and ConvNext, which adopted design principles from Vision Transformers (ViTs) (Jayanthi et al., 2026). The pure transformer architectures, when trained on large enough datasets, can perform as well as CNNs for image recognition tasks.

At the same time, self-supervised learning models like DINOv2 and contrastive vision-language models like CLIP have already trained feature extractors with impressive generalisation capabilities without the need for labelled pre-training data (Kaya & Eris, 2023). With this rapid development, there is a lack of systematic guidance to practitioners on which feature extractor is suitable for a specific task, with the availability of data and computational power. Comparisons which exist are normally performed within a single evaluation protocol or visual domain which makes it challenging to generalise. Also, there is a lack of a perfect comparison between modernised CNN architectures and ViT-based and self-supervised models (Vangipuram & Appusamy, 2025) in the frozen feature extraction paradigm for various downstream tasks. In this study, we fill this gap by systematically comparing seven pre-trained feature extractors (Geng et al., 2025) from different architecture of CNN, ViT, and self-supervised learning over five benchmark datasets from four different visual domains (Senanayake et al., 2023). Three complementary evaluation protocols were used: linear probing, k-nearest neighbour (k-NN) retrieval and 5-shot classification, each measuring a different aspect of the feature representations. The study is summarised by the following contributions:

- A comprehensive study of seven pre-trained feature extractors on five data sets and four visual domains that includes freezing all network parameters during testing.
- The results demonstrate that the modernised CNN ConvNeXt-B is better or comparable to ViT and self-supervised models in linear probe and few-shot classification tasks (Han et al., 2025).
- Results showing that the model ranking is dependent on the metric used, such as the k-NN retrieval ranking (Liu et al., 2026), with DINOv2-S being ranked above ConvNeXt-B, while being close to the same rank for accuracy-based metrics, highlighting the importance of multi-protocol evaluations.
- With a variety of image-text pre-training corpus, CLIP-ViT-B/32 is best suited for cross-domain generalisation to tasks such as texture recognition and satellite imagery (K. Niu et al., 2025).
- Model selection guidelines for practitioners, depending on task type, data availability and computer requirements.

Literature Review

The main challenges in machine learning and deep learning have been to find meaningful representation from raw data. Feature extraction methods are also the part of the classification pipelines and can be used to reduce the dimensionality of the input space and can extract features which can be learnt by the downstream models (Yan et al., 2025; Mufassirin & Amath, 2026). Many research has been done on this process in both classical and deep learning architecture, feature extraction architecture has been noted as one of the major aspects that affecting the model performance (Chu et al., 2025). Earlier studies with feature extraction techniques have focused on unsupervised dimensionality reduction techniques and combat curse of dimensionality in small and high dimensional data sets. This comparison of Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), Multidimensional Scaling, isometric mapping, Locally Linear Embedding (LLE), Laplacian Eigenmaps, Independent Component Analysis (ICA), and autoencoders has shown that they all perform better in different situations with different types of data (H. Niu et al., 2025). Datasets that containing linearly separable structures they are well suited and efficient to be used as a preprocessing step before classification, due to the low computation load since they are based on a projection of linear methods. If data are available on a curved manifold, other nonlinear manifold-based methods like ISOMAP and locally linear embedding would be more appropriate models (Miao & Xu, 2025), but they carry higher expensive and sensitive to neighbourhood size.

The linear and the neural network-based methods outperformed other spectral graph-based methods on the smaller datasets while the support vector machine (SVM) classifiers (Basthikodi et al., 2024) based on the spectral graph-based methods, especially Laplacian eigenmaps, yielded the most accurate classification on benchmark time-series datasets. This discovery is both useful and practically relevant since it is the geometric organisation of the feature space that is what guides the use of k-nearest neighbour retrieval as an evaluation protocol for the present study in addition to describing the quality of classification under a learned boundary. The most developed departure from hand-crafted to learned feature extraction has been in the representation of medical images (K. Niu et al., 2025), where the quality of the representation is more critical, and the complexity and variability of the medical imaging data is more apparent. A lot of work has been carried out to assess the effectiveness of the feature extraction of the conventional morphological features compared to the feature extraction from the deep convolutional neural network architectures such as the texture features (Al-Thelaya et al., 2023) derived from grey-level co-occurrence matrix, statistical moments, fractal dimension and wavelet coefficients.

The findings consistently show that deep learning features extracted from architectures such as ResNet, DenseNet, and Inception networks produce more discriminative embeddings for tissue classification, nuclei detection, and cancer grading tasks than handcrafted alternatives (Kaya & Eris, 2023). Among convolutional architectures, DenseNet-based feature extractors were found to capture more complex descriptive characteristics than earlier architectures such as VGG or Inception when evaluated on histopathological image retrieval tasks. ResNet-based models (Athisayamani et al., 2023) demonstrated strong transferability from ImageNet pre-training (Okazaki et al., 2024) to domain-shifted medical imaging domains, supporting their use across multiple downstream tasks including cancer detection, survival prediction, and image segmentation. These findings highlight that architectural depth, skip connections, and pre-training data scale are all influential factors in determining the quality of transferred features a finding that motivates the inclusion of ResNet-50, DenseNet-121, and ConvNeXt-B (Ramos, Casas, Romero, Rivas-Echeverría, & Morocho-Cayamcela, n.d., 2022) within the comparative framework of the present study.

Work on Alzheimer's disease diagnosis from brain MRI images has directly examined the interaction between pre-trained convolutional feature extractors and downstream classifier choice (AlSaeed & Omar, 2022). When ResNet-50 is used as a frozen feature extractor and its fully connected layer representations consistently achieves superior performance, reaching accuracy levels of 99% on the ADNI dataset. The SVM-based configuration achieves second-best performance at 92%, while random forest trails at 85.7%, demonstrating that the quality of the feature representation interacts significantly with the capacity of the downstream classifier (Youssef, Atef, 2025). This result confirms that linear probing training a logistic or softmax classifier on frozen features is a valid and widely used benchmark for assessing the raw quality of pre-trained representations, and that the linear separability of the feature space is a reliable proxy for overall transfer learning quality (Yildirim et al., 2023). The study also demonstrates that ResNet-50 features, when augmented with standard data augmentation and regularisation techniques, achieve accuracy levels comparable to or exceeding ensemble deep learning approaches that combine multiple pre-trained architectures, underscoring the practical significance of choosing a strong single feature extractor over complex multi-model pipelines. The feature extraction quality is shaped by a complex interplay of architectural design, pre-training data diversity, network depth, and the evaluation protocol used to measure representational quality. Existing studies largely evaluate individual extractors on single datasets or within narrow application domains, making cross-architecture, cross-domain comparisons rare.

The present study directly addresses this gap by providing the first systematic evaluation of seven pre-trained extractors spanning CNN, Vision Transformer (Sidiropoulos, Kiratsa, 2021), and self-supervised architectural families across five benchmark datasets and three complementary evaluation protocols linear probing, k-nearest neighbour retrieval, and few-shot classification thereby providing a comprehensive empirical foundation for evidence-based feature extractor selection in practical deep learning applications.

Materials and Methods

Feature Extractor Models

Seven pre-trained feature extractors were selected to represent the breadth of current architectural families (see Table 1). All models were loaded using publicly available pre-trained weights via the PyTorch Image Models(timm)library. The exact timm model identifiers used were: resnet50.a1_in1k, efficientnet_b4.ra2_in1k, densenet121.ra_in1k, convnext_base.fb_in22k_ft_in1k, vit_base_patch16_224.augreg2_in21k_ft_in1k, dinov2_vits14, and clip_vit_base_patch32 (via OpenCLIP). Pre-trained weights were obtained from the timm model hub (https://huggingface.co/timm) with default checkpoint versions as of the experiment date. Classification heads were removed by setting num_classes=0, yielding a global feature vector as output. All model weights were strictly frozen throughout the evaluation no fine-tuning of any kind was performed. Each model was configured with its recommended input resolution and channel-wise normalisation statistics, obtained automatically via timm's resolve_model_data_config() function, ensuring that each model processed images at its intended scale and normalisation. Notably, DINOv2-S requires 518×518-pixel inputs due to its patch size of 14 pixels, while EfficientNet-B4 operates at 380×380 pixels.

Table 1
Summary of pre-trained feature extractors evaluated in this study

Model	Family	Input	Feat. dim	Params	Pre-training data
ResNet-50	CNN	224×224	2048	25M	ImageNet-1k (supervised)
EfficientNet-B4	CNN	380×380	1792	19M	ImageNet-1k (supervised)
DenseNet-121	CNN	224×224	1024	8M	ImageNet-1k (supervised)
ConvNeXt-B	Modern CNN	224×224	1024	89M	ImageNet-21k (supervised)
ViT-B/16	Vision Transformer	224×224	768	86M	ImageNet-21k (supervised)
DINOv2-S	Vision Transformer	518×518	384	22M	LVD-142M (self-supervised)
CLIP-ViT-B/32	Vision Transformer	224×224	512	88M	LAION-2B (image-text pairs)

Datasets

Five benchmark datasets were selected to span a range of visual domains, class granularities, and degrees of domain shift from the ImageNet pre-training distribution and Collected from Kaggle (see Table 2):

- CIFAR-10 and CIFAR-100: General object detection benchmarks dataset with 10 and 100 classes respectively. Both contain 50,000 training and 10,000 test images at 32×32 pixels, resized to each model's required input size. CIFAR-10 used as a baseline with low class diversity; CIFAR-100 is a more challenging fine-grained discrimination task.
- Oxford-IIIT: Pets A fine-grained classification dataset containing 37 cat and dog breed classes. Approximately 3,680 images are available for training and 3,669 for testing. High visual similarity within the dataset makes this a challenging transfer learning benchmark.
- Describable Textures Dataset DTD: A texture recognition dataset with 47 categories (e.g., banded, bubbly, cracked, fibrous). Each split contains 1,880 images. DTD specifically targets the difference between texture-biased and shape-biased feature representations.
- EuroSAT: A satellite land-use classification dataset with 10 categories and 27,000 images at 64×64 pixels, resized to the model's required input size. As satellite imagery is markedly different in appearance from natural photographs, this dataset provides the strongest test of out-of-domain generalisation. explicitly randomly split the data set with 80:20 ratio using seed 42 for reproducibility.

Table 2
Benchmark datasets used for evaluation.

Dataset	Domain	Classes	Train	Test	Notes
CIFAR-10	General objects	10	50,000	10,000	Krizhevsky (2009)
CIFAR-100	General objects	100	50,000	10,000	Krizhevsky (2009)
Oxford Pets	Fine-grained	37	3,680	3,669	Parkhi et al. (2012)
DTD	Texture	47	1,880	1,880	Cimpoi et al. (2014)
EuroSAT	Satellite imagery	10	21,600	5,400	Helber et al. (2019); 80/20 split, seed=42

Evaluation Protocols

Three complementary evaluation protocols were used, each measuring a different aspect of the extracted feature representations:

***Linear Probing*:** A logistic regression classifier (regularization strength C=1.0, solver=lbfgs, maximum 1,000 iterations) was trained on StandardScaler-normalized features extracted from the frozen model. This protocol constitutes the primary benchmark used in the transfer learning literature and measures the overall linear discriminability of the feature space.

k-Nearest Neighbour Retrieval (k-NN): A cosine-similarity k-NN classifier (k=20) was applied to L2-normalised features. This protocol does not require any trained parameters and directly measures the geometric structure and natural class separability of the feature space without any decision boundary learning.

5-Shot Classification: A logistic regression classifier was trained only five randomly sampled examples per class. The experiment was repeated over 10 independent random episodes (seeds 42–51) and mean ± standard deviation accuracy is reported. This protocol measures low-data generalisation ability, which is critical for practical applications where labelled data acquisition is costly.

Implementation Details

All experiments were conducted on Kaggle computational notebooks equipped with two NVIDIA Tesla T4 GPU (15 GB VRAM). Figure 1 shows the system architecture of the proposed evaluation pipeline employed in this study, the Feature extraction was implemented in PyTorch 2.x, using torch.no_grad() context management and pin_memory=True DataLoaders for computational efficiency. A batch size of 128 was used for all models except DINOv2-S, for which batch size was reduced to 64 due to the larger input resolution. Each model was loaded once per session; upon completion of feature extraction for all five datasets, the model was deleted from GPU memory (torch.cuda.empty_cache(); gc.collect()) before loading the next model to prevent memory overflow. All extracted features were saved as NumPy binary arrays (.npy format) and reused across all evaluation tasks without re-running the forward pass. Standard deviations for 5-shot results ranged from 0.3% to 2.1% across all

model-dataset combinations. All random operations used global seed 42 for full reproducibility.

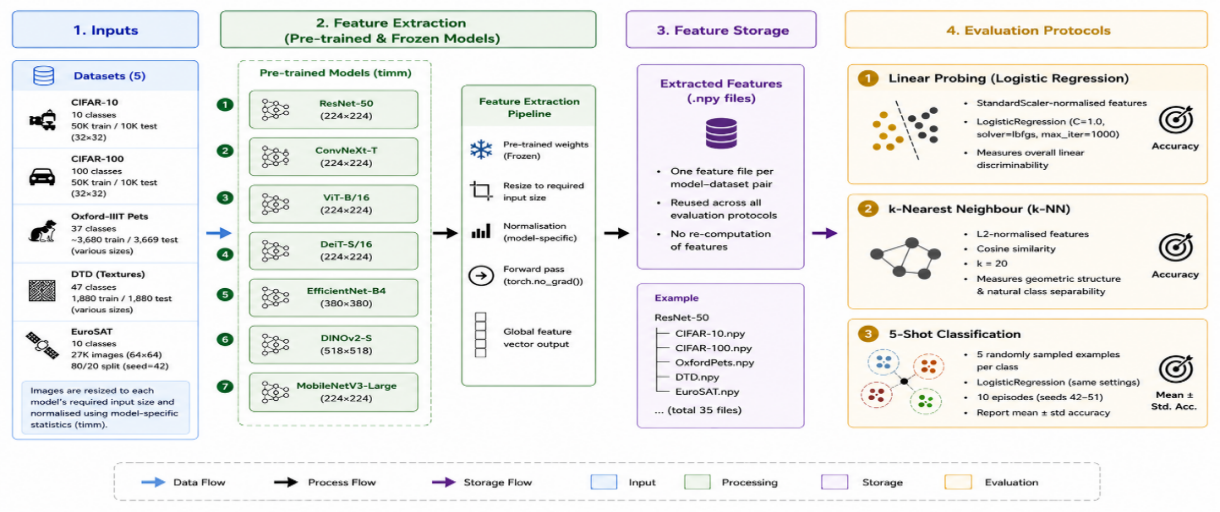

Figure 1. *System architecture of the proposed evaluation pipeline for comparing pre-trained feature extractor models.*

Results

Linear Probe Accuracy

Table 3 shows top-1 accuracy for all seven models on all five datasets under the linear probe protocol in the order of the mean accuracy. ConvNeXt-B has the best mean accuracy of 90.40%, followed closely by DINOv2-S (90.33%) and ViT-B/16 (90.15%). From these three top-performing models were separated by only 0.25% points, suggesting that modernised CNNs and transformer-based models can achieve good representations on large scale pre-training. CLIP-ViT-B/32 achieved the best overall score (88.47%) with the best individual score on DTD (79.15%). ViT-B/16 achieved the highest scores on CIFAR-10 (97.38%) and EuroSAT (96.83%), and DINOv2-S on Oxford-IIIT Pets (95.15%). ResNet-50 (83.43%) and DenseNet-121 (82.60%) ranked last by a consistent margin of approximately seven percentage points below the top cluster, confirming that older CNN architectures without modern design improvements produce substantially weaker transferable representations.

Table 3

Linear probe top-1 accuracy (%) for all models across all datasets, sorted by mean accuracy.

Model	CIFAR-10	CIFAR-100	DTD	EuroSAT	Oxford Pets	Mean
ConvNeXt-B	97.13	**86.62**	77.18	96.35	94.74	**90.40**
DINOv2-S	97.04	85.38	78.6s7	95.39	**95.15**	90.33
ViT-B/16	**97.38**	86.50	75.43	**96.83**	94.63	90.15
CLIP-ViT-B/32	95.39	81.15	**79.15**	96.52	**90.13**	88.47
EfficientNet-B4	94.68	78.64	70.27	95.65	94.03	86.65
ResNet-50	89.73	71.67	67.71	**95.35**	92.70	83.43
DenseNet-121	**89.24**	**71.38**	**66.44**	95.65	90.30	**82.60**

Note. Bold green indicates best per column and bold red indicate worst per column

K-NN Retrieval Accuracy

Table 4 reports k-NN (k=20, cosine similarity) accuracy on L2-normalised features. DINOv2-S ranked first with a mean of 87.64%, overtaking ConvNeXt-B (86.98%) under this training-free protocol. This reversal relative to linear probe rankings is notable: although both models score similarly under supervised probing, DINOv2-S features form more naturally separated class clusters in the embedding space, as evidenced by higher k-NN accuracy without any trained decision boundary. This finding is consistent with the self-supervised learning objective of DINOv2, which optimises for semantic cluster coherence rather than discriminative boundary placement. CLIP-ViT-B/32 achieved the highest individual k-NN score on EuroSAT (92.85%), further confirming cross-domain feature quality. The largest drop from linear probe to k-NN was observed for CLIP on Oxford-IIIT Pets (90.13% → 82.97%), suggesting that CLIP features for fine-grained tasks are better suited to learned boundaries than to proximity-based classification.

Table 4

k-NN (k=20, cosine similarity) accuracy (%) on L2-normalised features.

Model	CIFAR-10	CIFAR-100	DTD	EuroSAT	Oxford Pets	Mean
DINOv2-S	**97.59**	**85.29**	**70.43**	92.50	92.37	**87.64**
ConvNeXt-B	97.18	84.57	68.30	91.76	**93.08**	86.98
ViT-B/16	97.31	84.42	65.32	90.94	91.99	86.00
CLIP-ViT-B/32	95.24	79.38	**70.43**	**92.85**	**82.97**	84.17
EfficientNet-B4	94.00	74.46	63.03	**90.93**	92.83	83.05
ResNet-50	89.03	65.83	61.91	92.06	90.60	79.89
DenseNet-121	**87.87**	**65.07**	**60.21**	92.35	89.78	**79.06**

Note. Bold green indicates best per column and bold red indicate worst per column

Five-Shot Classification

Table 5 presents mean ± standard deviation accuracy across 10 random episodes of 5-shot classification. ConvNeXt-B achieved the highest mean 5-shot accuracy (80.23%), followed by ViT-B/16 (79.45%) and DINOv2-S (78.81%). Contrary to the common expectation that self-supervised models should dominate in low-data regimes, ConvNeXt-B and ViT-B/16 both trained with label supervision on ImageNet-21k outperformed DINOv2-S in overall mean accuracy, suggesting that the scale and label richness of ImageNet-21k supervised pre-training provides representations equally suitable for few-shot generalisation. CLIP-ViT-B/32 achieved the highest individual 5-shot score on EuroSAT (80.40%), consistent with the cross-domain transfer advantage observed under other protocols. The performance gap between the top cluster (ConvNeXt-B, ViT-B/16, DINOv2-S) and the bottom cluster (ResNet-50: 64.98%, DenseNet-121: 62.88%) widened substantially in the 5-shot setting relative to the linear probe setting, reaching approximately 17 percentage points, indicating that older CNN architectures are particularly ill-suited for low-data transfer scenarios.

Table 5
Five-shot mean accuracy (%) over 10 random episodes.

Model	CIFAR-10	CIFAR-100	DTD	EuroSAT	Oxford Pets	Mean
ConvNeXt-B	**93.36**	**77.18**	63.34	75.94	**91.32**	**80.23**
ViT-B/16	92.35	76.54	60.53	77.85	89.96	79.45
DINOv2-S	87.57	72.76	66.87	77.99	88.88	78.81
CLIP-ViT-B/32	89.44	68.17	**66.94**	**80.40**	77.93	76.58
EfficientNet-B4	84.24	59.83	54.01	75.28	89.62	72.60
ResNet-50	68.24	49.45	49.69	**71.20**	86.30	64.98
DenseNet-121	**66.47**	**46.46**	**48.76**	77.74	**74.95**	**62.88**

Note. Bold green indicates best per column and bold red indicate worst per column

Computational Cost

Extraction time varied substantially across architectures. Standard CNN models (ResNet-50, DenseNet-121, EfficientNet-B4) were the fastest due to their efficient convolutional operations. Transformer-based models required greater computation; DINOv2-S was the slowest extractor due to its high-resolution 518×518 pixel input requirement, which quadruples the token count relative to 224×224 ViT models. Approximate per-dataset feature extraction times on a single NVIDIA Tesla T4 GPU were as follows: ResNet-50 (~45 s), DenseNet-121 (~52 s), EfficientNet-B4 (~90 s), ConvNeXt-B (~110 s), ViT-B/16 (~120 s), CLIP-ViT-B/32 (~130 s), and DINOv2-S (~310 s). These figures reflect total wall-clock extraction time per dataset and are reported for indicative comparison; exact values may vary with batch size and dataset size. Figure 2 shows the accuracy heatmap of the all-evaluation tasks, and Figure 3 shows t-SNE projections of CIFAR-10 test features for all models.

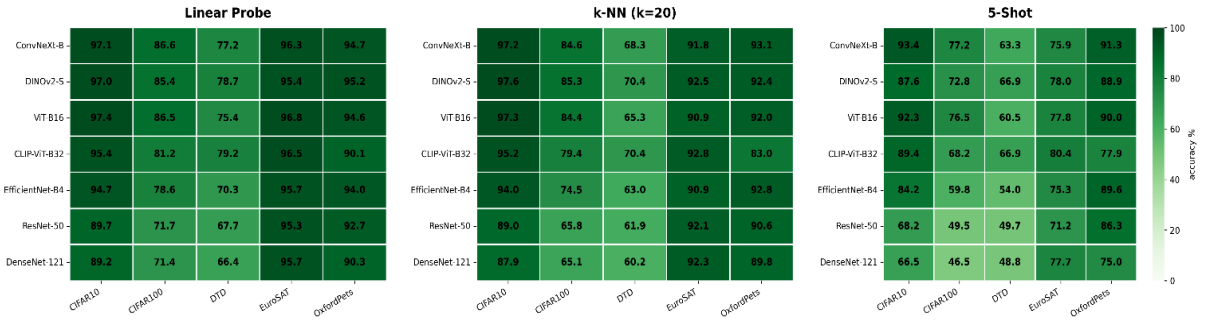

Figure 2. *Three-panel accuracy heatmap comparing all seven feature extractors across linear probe, k-NN retrieval, and 5-shot classification on five benchmark datasets. Darker green indicates higher accuracy.*

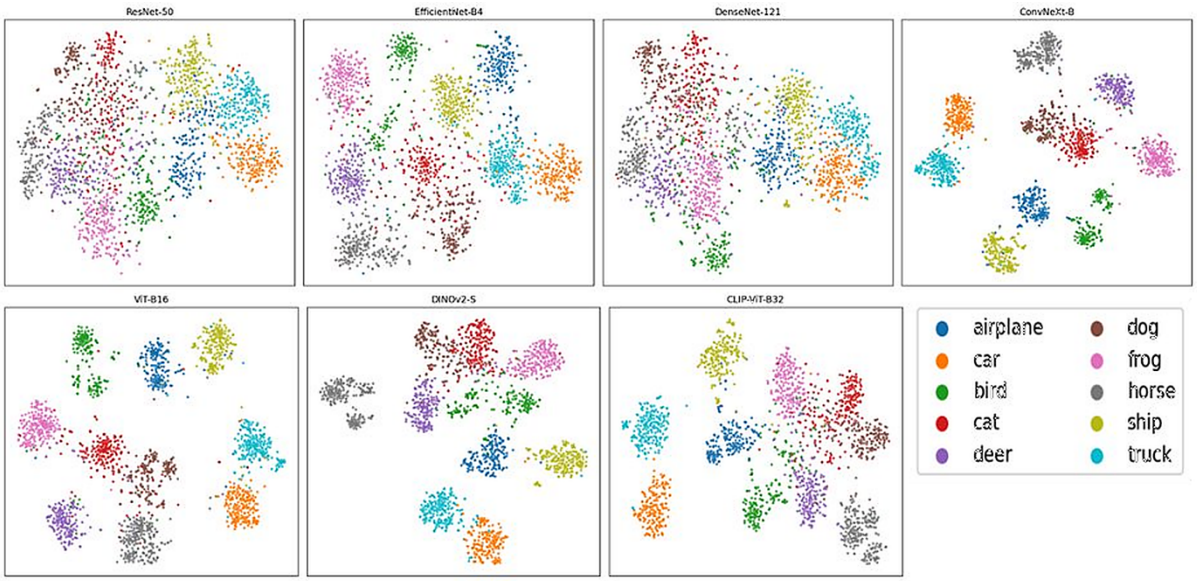

Figure 3. *t-SNE projections of CIFAR-10 test set features (2,000 samples) for all seven feature extractors. Each color represents one of the ten CIFAR-10 classes. Tighter, more separated clusters indicate better geometric organization of the feature space.*

Discussion

The results of this study show several findings that extend and, in some cases, challenge prevailing assumptions in the transfer learning literature. Perhaps the most notable finding of this study is that within the evaluated datasets and protocols, ConvNeXt-B achieved the best or near-best performance in most settings compared with the evaluated transformer and self-supervised models. This finding suggests that, under the specific pre-training and evaluation conditions of this study, a ViT-based models were not consistently superior to ConvNeXt-B across the evaluated tasks. The main reasons for the best results of ConvNeXt include: large-scale ImageNet-21k supervised pre-training using label rich supervision on 21,841 semantic categories; architectural modifications using the ViT-inspired design (depthwise convolution, inverted bottleneck, LayerNorm, and GELU activations) that help achieve higher representational quality than in standard CNNs; and keeping the spatial inductive bias of convolution that benefits classification-oriented tasks used in this study. This finding is consistent with prior work suggesting that carefully modernised CNN designs can match or outperform ViT baselines under comparable pre-training conditions, although these observations are limited to the datasets, model variants, and evaluation protocols considered in this study, and further investigation is needed to assess their generalisability to other tasks and model scales.

The change in models ranking for linear probe and k-NN protocols gives an important insight into the nature of the various feature spaces. DINOv2-S outperforms ConvNeXt-B in terms of k-NN retrieval for almost all datasets while performing comparably with linear probe accuracy. The performance of CLIP-ViT-B/32 on texture-dominant and cross-domain tasks provides further evidence for the value of pre-training data diversity. CLIP leads on DTD under both linear probe (79.15%) and 5-shot classification (66.94%), and achieves the highest 5-shot score on EuroSAT (80.40%). These results are mechanistically consistent with CLIP's pre-training corpus: internet image-text pairs naturally contain diverse texture descriptions and include satellite and remote sensing imagery alongside descriptive captions, exposing the model to a far broader distribution of visual concepts than ImageNet-based models. This finding supports the position of that language-grounded supervision produces semantically richer visual representations, particularly for distribution-shifted domains. The consistent

underperformance of ResNet-50 and DenseNet-121 lagging approximately seven percentage points below the top cluster on linear probe and 17 percentage points on 5-shot classification confirms that older CNN architectures have been superseded as general-purpose feature extractors.

These models lack the architectural refinements, pre-training data scale, and optimisation improvements that characterise modern alternatives. Their retention in practice is justifiable only in settings with severe computational constraints or when inference latency is the primary concern, given their superior speed relative to transformer-based models. This study has several limitations. First, evaluation is restricted to image classification tasks; the relative merits of these feature extractors on dense prediction tasks (object detection, semantic segmentation) may differ substantially. Second, only one model variant of the architectural families has been assessed, larger model variants (ViT-L, DINOv2-L, ConvNeXt-XL) may give different relative rankings, especially for self-supervised models, which may benefit more from scale. Thirdly, the 80/20 split was random and there was no fixed official test partition in the EuroSAT dataset so that it is possible that this could have an impact on comparability with other published results. Further studies could be conducted with larger model variations, video understanding tasks, and medical imaging applications, and should consider how partial fine-tuning methods affect performance compared to fully freezing the extraction over various amounts of labelled data.

Conclusion

This study addresses a fundamental challenge in transfer learning and deep learning applications: determining which pre-trained feature extractor yields the most effective feature representations across diverse tasks and visual domains. For this purpose, seven models from different families (convolutional, transformer, self-supervised) were evaluated on five benchmark datasets, following three evaluation protocols. The results did not identify a universally superior model across all evaluation settings. The overall linear probe accuracy of 90.40% achieved by ConvNeXt-B, a modernised convolutional network trained on ImageNet-21k, indicates that, under the evaluated conditions, a well-designed CNN can perform competitively with the evaluated Vision Transformer models on the image classification tasks considered in this study. In the case of DINOv2-S, the self-supervised pre-training approach, the feature geometry was the cleanest, organized in a fundamentally different manner from the label, and more geometrically coherent than that of other methods. The results of the few-shot protocol in texture recognition task and satellite imagery shows that CLIP-ViT-B/32 has a clear advantage, most likely because of the diversity and volume of the image-text pre-training data, but not necessarily because of its architecture.

The other two older architectures, DenseNet-121 and ResNet-50, consistently achieved lower performance than the top-performing models across the evaluated tasks, suggesting that architectural design and pre-training scale may influence transfer learning performance within the scope of the classification tasks examined in this study. Altogether, these observations suggest that in the specific downstream task, all of these show a preference for choosing model selection, ConvNeXt-B for high accuracy classification (when some labelled downstream data exists), DINOv2-S for retrieval or clustering (when no labelled downstream data exists), and CLIP-ViT-B/32 for a target domain other than natural photography. A larger model size, video and medical imaging applications, and cases where some fine-tuning (non-complete freezing of the extraction) is permitted should be included in future studies; the relative ranks obtained here may be quite different if some fine-tuning (not complete freezing of the extraction) can be made somehow.

References

AlSaeed, D., & Omar, S. F. (2022). Brain MRI Analysis for Alzheimer's Disease Diagnosis Using CNN-Based Feature Extraction and Machine Learning. *Sensors*, *22*(8). https://doi.org/10.3390/s22082911

Al-Thelaya, K., Gilal, N. U., Alzubaidi, M., Majeed, F., Agus, M., Schneider, J., & Househ, M. (2023). Applications of discriminative and deep learning feature extraction methods for whole slide image analysis: A survey. In *Journal of Pathology Informatics* (Vol. 14). Elsevier B.V. https://doi.org/10.1016/j.jpi.2023.100335

Athisayamani, S., Antonyswamy, R. S., Sarveshwaran, V., Almeshari, M., Alzamil, Y., & Ravi, V. (2023). Feature Extraction Using a Residual Deep Convolutional Neural Network (ResNet-152) and Optimized Feature Dimension Reduction for MRI Brain Tumor Classification. *Diagnostics*, *13*(4). https://doi.org/10.3390/diagnostics13040668

Basthikodi, M., Chaithrashree, M., Ahamed Shafeeq, B. M., & Gurpur, A. P. (2024). Enhancing multiclass brain tumor diagnosis using SVM and innovative feature extraction techniques. *Scientific Reports*, *14*. https://doi.org/10.1038/s41598-024-77243-7

Chu, W., Wen, H., Liu, H., Zhang, X., & Guo, J. (2025). Fusion of EEG feature extraction and CNN-MSTA transformer emotion recognition classification model. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-28470-z

Geng, X., Hu, W., Liang, Q., Chen, C., Zhang, X., Yu, P., & Bao, J. (2025). An improved feature extraction algorithm for EEG-based driving fatigue recognition. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-18554-1

Han, Y., Han, W., Li, A., & Li, S. (2025). Cyberattack event and arguments extraction based on feature interaction and few-shot learning. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-15138-x

Jayanthi, S., Kaur, I., Laxmi Lydia, E., Kumar, K. V., Joshi, G. P., & Cho, W. (2026). Transformer-assisted convolutional feature extraction with deep representation learning models for lung and colon cancer diagnosis using histopathological images. *Scientific Reports*, *16*. https://doi.org/10.1038/s41598-025-34160-7

Kaya, M., & Eris, M. (2023). D3SENet: A hybrid deep feature extraction network for Covid-19 classification using chest X-ray images. *Biomedical Signal Processing and Control*, *82*. https://doi.org/10.1016/j.bspc.2022.104559

Li, Y., Zhang, L., Chen, L., & Ma, Y. (2025). Superpixel guided spectral-spatial feature extraction and weighted feature fusion for hyperspectral image classification with limited training samples. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-87030-7

Liu, J., Ahmad, F. A., Samsudin, K., Hashim, F., & Kadir, M. Z. A. A. (2026). Deep residual networks with convolutional feature extraction for short-term load forecasting. *Scientific Reports*, *16*. https://doi.org/10.1038/s41598-026-35410-y

Miao, B., & Xu, C. (2025). Aspect-level multimodal sentiment analysis model based on multi-scale feature extraction. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-16051-z

Mufassirin, M. M., & Amath, A. A. S. (2026). Advanced Feature Selection and Extraction Techniques for Omics Data Analysis: Applications in Multi-Omics Integration. In Feature Selection and Feature Extraction on Omics Data (pp. 51-72). Chapman and Hall/CRC.

Niu, H., McCallum, G. B., Chang, A. B., Khan, K., & Azam, S. (2025). Exploring unsupervised feature extraction algorithms: tackling high dimensionality in small datasets. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-07725-9

Niu, K., Han, J., & Cai, J. (2025). CFM-UNet: coupling local and global feature extraction networks for medical image segmentation. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-92010-y

Okazaki, S., Mine, Y., Yoshimi, Y., Iwamoto, Y., Ito, S., Peng, T.-Y., Nishimura, T., Suehiro, T., Koizumi, Y., Nomura, R., Tanimoto, K., Kakimoto, N., & Murayama, T. (2024). RadImageNet and ImageNet as Datasets for Transfer Learning in the Assessment of Dental Radiographs: A Comparative Study. *Journal of Imaging Informatics in Medicine*, *38*. https://doi.org/10.1007/s10278-024-01204-9

Ramos, L., Casas, E., Romero, C., Rivas-Echeverría, F., & Morocho-Cayamcela, M. E. (n.d.). (2024). A Study of ConvNeXt Architectures for Enhanced Image Captioning. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3356551

Senanayake, D., Seneviratne, S., Imani, M., Harijanto, C., Sales, M., Lee, P., Duque, G., & Ackland, D. C. (2023). Classification of Fracture Risk in Fallers Using Dual-Energy X-Ray Absorptiometry (DXA) Images and Deep Learning-Based Feature Extraction. *JBMR Plus*, *7*(12).

Sidiropoulos, G. K., Kiratsa, P., Chatzipetrou, P., & Papakostas, G. A. (2021). Feature extraction for finger-vein-based identity recognition. In *Journal of Imaging* (Vol. 7, Number 5). MDPI AG. https://doi.org/10.3390/jimaging7050089

Vangipuram, S. K., & Appusamy, R. (2025). A Novel Image Feature Extraction Based Machine Learning Approach for Disease Detection from Chest X-Ray Images. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *7*(1), 56–79. https://doi.org/10.35882/jeeemi.v7i1.529

Yan, Q., Zhang, S., Chen, X., & Zheng, Z. (2025). Multiscale superpixel depth feature extraction for hyperspectral image classification. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-90228-4

Yildirim, B., Ulkir, O., Kaya, M., & Singh, A. K. (2023). *Trends in EEG signal feature extraction applications*. https://doi.org/10.3389/frai.2022.1072801

Youssef, D., Atef, H., Gamal, S., El-Azab, J., & Ismail, T. (2025). Early Breast Cancer Prediction Using Thermal Images and Hybrid Feature Extraction-Based System. *IEEE Access*, *13:29327-29339*. https://doi.org/10.1109/access.2025.3541051