Plant Systematics and Evolution (2026) 312:30
https://doi.org/10.1007/s00606-026-02005-z

RESEARCH

Balancing accuracy and efficiency in fine-grained image classification: a systematic comparison of CNN, vision transformer, and hybrid architectures

Demir Mesut Öztaş[1] · Hakan Salehi[1] · Özge Demir[2] · Koray Acici[3] · Mehmet Serdar Güzel[4] · Emre Cem Eraslan[5] · Mustafa Sevindik[5] · Ilgaz Akata[6] · Fatih Ekinci[7]

Received: 28 January 2026 / Accepted: 24 April 2026
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2026

Abstract
Fine-grained image classification remains a challenging problem in computer vision due to high inter-class visual similarity and pronounced intra-class variability. In this study, a systematic and architecture-centric comparison of convolutional neural networks (CNNs), Vision Transformer (ViT) models, and hybrid architectures is presented using the Oxford 102 Flowers dataset, which comprises 8189 images across 102 visually similar categories. Rather than proposing a new architecture or training strategy, this work focuses on a controlled and comparative analysis of widely adopted deep learning paradigms under a unified experimental framework. All models are evaluated under a strictly controlled experimental protocol with fixed data splits, identical preprocessing, and unified optimization settings to reduce experimental confounding factors and enable architecture-level analysis. The experimental results show that large-scale Vision Transformers achieve the highest absolute classification performance within this benchmark setting, with ViT-L/16 attaining an accuracy of 99.37% and an F1-score of 99.27%. However, this performance is accompanied by substantial computational cost, including 303.4 million parameters and an inference time of 27.4 ms per image. In contrast, hybrid architectures exhibit a more favorable accuracy–efficiency trade-off under the same experimental conditions. Notably, ConvNeXt-V2-Tiny achieves 98.13% accuracy and a 98.05% F1-score while requiring only 27.9 million parameters and an inference time below 1 ms, achieving competitive performance with significantly lower computational complexity. Conventional CNN architectures demonstrate comparatively lower performance within this specific transfer learning configuration, with average accuracy remaining below 90%, suggesting potential limitations in capturing highly subtle inter-class variations under the evaluated setup rather than indicating general architectural inadequacy. The findings indicate that the highest classification accuracy does not necessarily correspond to the most practically efficient solution, particularly in deployment-constrained environments. Importantly, the conclusions of this study are restricted to the Oxford 102 Flowers dataset under transfer learning and standardized input resolution, and should not be interpreted as universal claims of architectural superiority. By jointly analyzing accuracy, model size, and inference latency, this work provides a structured and decision-oriented assessment of architectural trade-offs for fine-grained image classification. The results offer practical insights for selecting appropriate deep learning architectures under varying computational constraints and contribute to a more nuanced understanding of how convolutional, transformer-based, and hybrid representations behave within a controlled fine-grained benchmark setting.

Keywords Computational efficiency · Convolutional neural networks · Fine-grained image classification · Hybrid deep learning architectures · Vision Transformers

Introduction

Fine-grained image classification constitutes one of the most challenging problem settings in computer vision, as it requires the discrimination of categories that exhibit high inter-class visual similarity while simultaneously accommodating substantial intra-class variability (Wei et al. 2021;

Handling Editor: Karol Marhold.

Extended author information available on the last page of the article

Published online: 07 June 2026

Kumru et al. 2025a, 2025b). Such characteristics impose strict requirements on representation learning, as successful models must capture subtle morphological, textural, and color-based cues rather than relying solely on coarse global shapes (Chen et al. 2025; Struniawski et al. 2026). To ensure methodological rigor and reproducibility, the official training, validation and test split of the dataset is strictly preserved, and all images are standardized to a fixed spatial resolution (Maier-Hein et al. 2024; Korkmaz et al. 2025). Within artificial intelligence, particularly deep learning serves as a dominant methodological paradigm, enabling data-driven extraction of hierarchical visual representations that are otherwise difficult to encode using hand-crafted features (Manakitsa et al. 2024; Luo et al. 2025).

Recent advances in deep learning have introduced diverse architectural paradigms for visual recognition, each characterized by distinct inductive biases and representational mechanisms (Castellano and Vessio 2021; Zhang et al. 2023). Three major classes of artificial intelligence models are systematically evaluated in this study: convolutional neural networks (CNNs), pure Vision Transformer (ViT) architectures, and hybrid models that integrate convolutional operations with attention-based mechanisms (Takahashi et al. 2024; Aamir et al. 2025; Alomar et al. 2025). Classical CNNs, such as VGG and ResNet families, primarily rely on localized receptive fields and hierarchical feature aggregation, making them effective for capturing fine textures but limited in explicit global context modeling (Li et al. 2024; Younesi et al. 2024). Vision Transformers, in contrast, employ self-attention mechanisms to model long-range dependencies across entire images, offering enhanced global representation capacity at the cost of increased computational complexity (Khan et al. 2022; Hassija et al. 2025). Hybrid architectures aim to bridge these paradigms by combining convolutional inductive biases with transformer-style attention, thereby seeking a balance between accuracy and computational efficiency (Wu et al. 2025; Zhang et al. 2025). All models are trained under a unified experimental protocol to reduce confounding factors and to examine architectural design choices under controlled conditions (Khan et al. 2022; Hassija et al. 2025; Wu et al. 2025; Zhang et al. 2025).

Beyond architectural innovation, recent literature (2023–2025) has increasingly emphasized comparative benchmarking, deployment constraints, and robustness-aware evaluation frameworks. In addition, recent studies have increasingly emphasized the importance of positioning comparative deep learning analyses within deployment-oriented and constraint-aware frameworks, rather than focusing solely on peak benchmark accuracy. In this context, the present study is explicitly framed as a controlled, architecture-centric benchmarking effort that aims to provide decision-oriented insights into accuracy–efficiency trade-offs, rather than proposing a novel model or training strategy. Contemporary reviews and cross-domain analyses argue that peak benchmark accuracy alone is insufficient for assessing practical model utility, advocating instead for multi-dimensional evaluation criteria incorporating efficiency, stability, and interpretability considerations (Khan et al. 2022; Maier-Hein et al. 2024; Takahashi et al. 2024). In particular, hybrid architectures and co-design strategies have been explored as mechanisms to reconcile global contextual modeling with computational feasibility (Dai et al. 2021; Woo et al. 2023). While many of these recent studies do not focus specifically on botanical fine-grained datasets, they provide a methodological foundation for positioning architecture comparisons within broader deployment-oriented and constraint-aware perspectives.

Unlike studies that propose new architectures or optimization strategies, the present work does not aim to introduce methodological innovations. Instead, it provides a controlled and comparative evaluation of representative CNN, ViT, and hybrid models under identical experimental settings in order to analyze accuracy–efficiency trade-offs in a fine-grained benchmark scenario. The importance of this study lies in its architecture-centric and decision-oriented perspective, which extends beyond reporting peak accuracy to systematically analyze how representational design principles relate to computational cost. By benchmarking CNN, ViT, and hybrid architectures under identical data splits, preprocessing pipelines, and optimization strategies, the study offers a transparent comparison framework in which performance differences can be interpreted in relation to architectural characteristics rather than experimental variability.

Importantly, the conclusions drawn in this work are restricted to the Oxford 102 Flowers dataset under transfer learning and standardized input resolution. The study does not claim universal architectural superiority across all fine-grained datasets; rather, it seeks to provide structured empirical evidence within a controlled benchmark context. The findings indicate that, within this setting, transformer-based models reach the highest absolute accuracy values, while certain hybrid architectures achieve competitive performance with substantially lower computational requirements.

Accordingly, the contribution of this work is analytical rather than architectural, focusing on systematic trade-off evaluation to inform deployment-aware model selection in fine-grained image classification. Beyond its immediate empirical findings, this controlled isolation protocol may serve as a reproducible reference framework for future fine-grained architectural evaluations, enabling subsequent studies to compare emerging models under harmonized and deployment-aware experimental conditions. Recent studies (2023–2025) emphasize efficiency-aware evaluation, highlighting cost–performance trade-offs, scaling behavior, and deployment constraints. Hybrid architectures and co-design strategies have been proposed to balance accuracy and

computational cost. However, many studies rely on heterogeneous settings, limiting direct comparability. Therefore, a controlled, architecture-centric evaluation remains necessary to isolate design effects and clarify accuracy–efficiency trade-offs (Rusia and Singh 2023; Alsalihi and Sztahó 2024; Raposo 2024; Ferreira and Couto 2025a, b, c; Ferreira et al. 2025b).

Materials and methods

This study aims to conduct a systematic and architecture-centric evaluation of fine-grained image classification frameworks by examining their representational expressiveness, generalization capability, and computational efficiency. In this context, convolutional neural networks (CNNs), pure Vision Transformer (ViT) architectures, and hybrid models that combine convolutional inductive biases with transformer-based global self-attention mechanisms are comparatively analyzed under strictly controlled and identical experimental settings. By enforcing a unified data partitioning scheme and consistent training protocols across all models, the study isolates the effect of architectural design choices, ensuring that the observed performance variations can be attributed primarily to differences in model structure rather than confounding factors such as dataset splits or optimization procedures.

Dataset description and experimental protocol

All experiments conducted in this study were carried out on the Oxford 102 Flowers dataset (Nilsback and Zisserman 2008), which is widely adopted in the literature as a benchmark for fine-grained image classification tasks. The dataset comprises a total of 8189 images belonging to 102 distinct flower categories, with the number of samples per class ranging from 40 to 258 (Nilsback and Zisserman 2008). This uneven distribution results in a pronounced class imbalance and a high degree of visual similarity across categories. Such characteristics impose significant challenges on the classification process, requiring models to capture not only global shape information but also fine-grained textural and morphological discriminative cues (Nilsback and Zisserman 2008). Representative sample images from the dataset are illustrated in Fig. 1.

The Oxford 102 Flowers dataset contains 8189 images across 102 classes, with the number of images per class ranging from 40 to 258, indicating a non-uniform class distribution. This corresponds to an imbalance ratio of $258/40 \approx 6.45$ between the largest and smallest classes, which may affect per-class performance and error distribution in fine-grained recognition.

To ensure strict reproducibility and comparability, the official split provided with the dataset was preserved. Specifically, the official partition includes 1020 training images (10 per class), 1020 validation images (10 per class), and



Fig. 1 Example flower images used in this study: **a** *Passiflora caerulea* (passionflower), **b** *Nymphaea* sp. (water lily), **c** *Cyclamen persicum* (cyclamen), **d** *Erysimum cheiri* (wallflower)

6149 test images (remaining samples), enabling consistent benchmarking across architectures without split-induced variance. In addition to reporting aggregate accuracy and macro-F1, we therefore interpret results with explicit attention to class imbalance and class-level error concentration, as these effects can be masked by high overall accuracy in near-saturated transfer learning regimes.

It is important to note that Oxford 102 Flowers is considered a relatively mature and near-saturated benchmark under transfer learning settings. Due to its moderate scale and the widespread availability of ImageNet-pretrained models, very high classification accuracies have been reported in recent literature. Therefore, the dataset provides a controlled fine-grained evaluation environment but does not necessarily represent the full complexity of more challenging contemporary fine-grained benchmarks.

To ensure methodologically consistent experimental comparisons, the official training–validation–test split provided with the dataset was strictly preserved throughout all experiments (Nilsback and Zisserman 2008). This design choice aims to reduce methodological ambiguities that may arise from employing different data partitioning strategies across models. By adopting a fixed split configuration, the study strengthens the interpretability of architectural comparisons, while acknowledging that a fixed split alone does not eliminate all potential sources of experimental bias. Consequently, all models were evaluated on identical samples, enabling direct, transparent, and reproducible comparisons (Nilsback and Zisserman 2008).

To establish a comparable experimental setting across different architectural paradigms, all images were rescaled to a unified spatial resolution of 224×224 pixels (Nilsback and Zisserman 2008). This input standardization provides a common representational basis for convolutional neural networks, Vision Transformer-based models, and hybrid architectures, thereby allowing direct comparison of model outputs. During training, data augmentation techniques such as random resizing and cropping, horizontal flipping, and color perturbations were employed. These augmentations are intended to mitigate the risk of overfitting, particularly prevalent in fine-grained classification scenarios, and to promote the learning of more generalizable representations that are robust to intra-class variability.

In contrast, only deterministic preprocessing steps were applied during the validation and test phases. This strategy prevents performance measurements from being influenced by stochastic data transformations and enhances the stability of the evaluation process. All images were normalized using the mean and standard deviation statistics of the ImageNet dataset (Nilsback and Zisserman 2008). Such normalization facilitates the effective utilization of pretrained network weights and provides a more balanced initialization point across different architectures in the transfer learning setting.

Taken together, the structural characteristics of the dataset, the fixed experimental split strategy, and the consistent input standardization across architectures constitute a controlled experimental framework for architecture-centric analysis. However, conclusions drawn from this benchmark should be interpreted within the scope of Oxford 102 Flowers under transfer learning and standardized resolution, and should not be generalized to more complex fine-grained datasets such as CUB-200, FGVC-Aircraft, or large-scale ecological benchmarks without further empirical validation. As a result, the methodological coherence of the performance evaluations and architectural comparisons presented in the subsequent sections is directly supported, while their external validity remains bounded by the characteristics of this dataset.

Architecture selection and model descriptions

The models evaluated in this study are categorized under three fundamental paradigms based on their architectural design principles and representation learning mechanisms: convolutional neural networks (CNNs), pure Vision Transformer (ViT) architectures, and hybrid approaches. This categorization is not limited to comparing the final classification accuracy of different model families; rather, it enables a systematic analysis of how the balance between local feature extraction and global context modeling influences performance in fine-grained image classification tasks (He et al. 2016).

Among the CNN-based architectures, ResNet-50 represents a classical approach that addresses the vanishing gradient problem encountered in deep networks through residual connections (He et al. 2016). By introducing identity mappings between layers, this architecture enables stable training of deeper networks and has long served as a strong and reliable reference model in a wide range of visual recognition tasks. However, the representational capacity of ResNet-50 is primarily constrained by convolutional operations with local receptive fields, while global contextual information is modeled only indirectly and progressively through network depth. In this respect, ResNet-50 provides a robust baseline that exemplifies the conventional CNN paradigm in fine-grained classification scenarios (Simonyan and Zisserman 2015; He et al. 2016; Tan and Le 2019; Dosovitskiy et al. 2021).

The VGG-16 architecture, composed of sequential 3×3 convolutional layers arranged in a homogeneous and deep structure, has played a historically significant role in analyzing the representational capacity of convolutional networks (Simonyan and Zisserman 2015). While its simple and regular architectural design facilitates interpretability and analytical clarity, its high parameter count and substantial computational cost render VGG-16 less efficient compared

to more modern architectures. In this study, VGG-16 is included to evaluate the impact of architectural simplicity and parameter density on fine-grained classification performance (Simonyan and Zisserman 2015).

EfficientNet-B0 is based on a compound scaling strategy that jointly balances network depth, width, and input resolution (Tan and Le 2019). Designed with parameter efficiency as a primary objective, this architecture aims to deliver competitive performance under relatively low computational cost. Within the context of fine-grained classification, EfficientNet-B0 is incorporated to investigate to what extent local feature extraction capabilities can be enhanced through a more balanced and efficient network design (Tan and Le 2019).

Among pure Vision Transformer architectures, ViT-B/16 operates by partitioning an image into fixed size patches, each of which is embedded into a linear representation and processed via multi-head self-attention mechanisms (Dosovitskiy et al. 2021). By relying on global context modeling rather than convolutional inductive biases, this approach offers advantages in capturing long-range dependencies across the entire image. In this study, ViT-B/16 is evaluated as a medium-scale transformer architecture (Dosovitskiy et al. 2021).

ViT-L/16 extends this paradigm by adopting a deeper and wider structure than ViT-B/16, thereby providing increased representational capacity (Dosovitskiy et al. 2021). The expanded model capacity enables more effective learning of discriminative global patterns, which is particularly relevant in fine-grained classification problems. This model is included to examine the trade-off between potential accuracy gains and the associated increases in model size and computational cost (Dosovitskiy et al. 2021).

The DeiT-Base model represents a variant of the Vision Transformer architecture designed to achieve more efficient training (Touvron et al. 2021). Through the use of knowledge distillation-based training strategies, DeiT aims to attain competitive performance with comparatively less training data (Touvron et al. 2021). In this regard, DeiT-Base serves as an appropriate representative for analyzing the data efficiency of pure transformer architectures in fine-grained classification settings (Touvron et al. 2021).

Within the category of hybrid architectures, ConvNeXt-V2-Tiny offers a modernized reinterpretation of classical convolutional structures by adopting design elements inspired by transformer architectures, including layer normalization schemes, activation functions, and block-level design principles (Woo et al. 2023). While preserving local feature extraction capabilities, this architecture seeks to learn more globally coherent and balanced representations, making it a compelling example for studying the trade-off between performance and computational efficiency (Woo et al. 2023).

Swin-Tiny is a hierarchical Vision Transformer architecture that employs shifted window-based attention mechanisms to efficiently model both local and semi-global contextual information (Liu et al. 2021). This strategy mitigates the high computational cost associated with full global attention while enabling the learning of broader contextual representations compared to conventional convolutional networks (Liu et al. 2021).

MaxViT-Tiny is a hybrid architecture that integrates convolutional blocks with both local and global attention mechanisms within a unified architectural framework (Tu et al. 2022). By facilitating the learning of multi-scale representations, this design is particularly well-suited for fine-grained classification problems characterized by complex visual patterns (Tu et al. 2022).

Finally, CoAtNet-0 introduces an architecture that sequentially and progressively combines convolutional layers with attention-based layers (Dai et al. 2021). By focusing on local feature extraction in early stages and global context modeling in deeper layers, this design establishes a structural bridge between CNN and transformer-based paradigms (Dai et al. 2021).

This architectural diversity enables a comprehensive analysis of how different design principles balance representational capacity, generalization ability, and computational efficiency in fine-grained image classification tasks. Consequently, the study aims not only to identify which model achieves higher accuracy, but also to elucidate under which conditions specific architectural approaches are more appropriate (Dai et al. 2021; Liu et al. 2021; Tu et al. 2022).

All models were trained using a transfer learning paradigm with weights pretrained on ImageNet-1 K or ImageNet-21 K. This choice is methodologically critical given the relatively limited number of samples in the Oxford 102 Flowers dataset (Nilsback and Zisserman 2008) and the high degree of visual similarity across classes. Leveraging pretrained representations allows the models to focus directly on discriminative morphological and textural cues specific to the fine-grained classification task, rather than learning generic visual features from scratch. This approach contributes to more stable training dynamics, particularly for high-capacity architectures, and supports improved generalization performance (Dai et al. 2021; Liu et al. 2021; Touvron et al. 2021; Tu et al. 2022; Woo et al. 2023).

For optimization, the AdamW algorithm was employed consistently across all architectures. By decoupling weight decay from the gradient update mechanism, AdamW serves as an effective regularizer, especially for deep and high-parameter models, helping to control the risk of overfitting (Loshchilov and Hutter 2019). The use of a unified optimization strategy ensures a consistent experimental framework, allowing observed performance differences to be attributed

primarily to architectural design characteristics rather than optimization artifacts (Loshchilov and Hutter 2019).

Learning rates were adjusted according to the structural properties of the architectural paradigms. Higher initial learning rates were adopted for CNN-based models, whereas lower learning rates were used for Vision Transformer and hybrid architectures (Dai et al. 2021; Dosovitskiy et al. 2021; Liu et al. 2021; Touvron et al. 2021; Tu et al. 2022; Woo et al. 2023). While this approach ensures stable optimization for each architecture, it introduces a potential trade-off between optimization fairness and strict comparability. The use of architecture-dependent learning rates may influence performance outcomes, and therefore should be considered when interpreting comparative results. Convolutional networks, due to their strong inductive biases derived from local receptive fields, can be optimized stably even under relatively aggressive learning rates. In contrast, Vision Transformers and hybrid models relying partially or fully on attention mechanisms exhibit more sensitive gradient dynamics and may suffer from optimization instability under high learning rates (Loshchilov and Hutter 2019; Dosovitskiy et al. 2021). Accordingly, architecture-aware learning rate selection was intentionally applied to ensure fair and balanced comparisons across models.

The training process was monitored using an early stopping mechanism. This strategy allows each architecture to terminate training at an optimal point without excessive fitting, in accordance with its representational capacity. Particularly for high-capacity Vision Transformer and hybrid models, early stopping plays a complementary role in preserving generalization performance (Loshchilov and Hutter 2019).

The combination of transfer learning, architecture-sensitive learning rate adjustment, and a unified optimization framework enables the CNN, Vision Transformer, and hybrid models examined in this study to be evaluated within a fair, comparable, and architecture-focused experimental setting. As a result, the observed performance differences can be directly associated with representational mechanisms inherent to the architectures, thereby ensuring methodological consistency (Dai et al. 2021; Liu et al. 2021; Touvron et al. 2021; Tu et al. 2022; Woo et al. 2023).

The performance of the evaluated models was quantitatively assessed on the test set using accuracy and F1-score metrics. While these metrics provide a global performance overview, they do not fully capture class-level behavior, error distribution, or performance under class imbalance conditions. These measures are intended to reflect not only overall classification performance but also class-level balance, which is critical in fine-grained classification problems (Sokolova and Lapalme 2009; Chicco and Jurman 2020). Given the high inter-class visual similarity and variability in the number of samples per class in the Oxford 102 Flowers dataset (Nilsback and Zisserman 2008), accuracy alone may be insufficient to fully characterize model behavior. Therefore, the F1-score was included as a complementary metric to provide a more balanced evaluation (Sokolova and Lapalme 2009; Chicco and Jurman 2020). However, more detailed evaluation metrics such as per-class precision/recall, confusion matrices, and balanced accuracy could provide additional insights into model performance across individual categories.

Accuracy is defined as the ratio of correctly classified samples to the total number of samples in the test set and is computed as follows:

$$Accuracy = \frac{TP + TN}{FP + FN} \tag{1}$$

Here, TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. In the context of multi-class classification, accuracy is computed as the ratio of the total number of correctly predicted samples across all classes to the total number of samples (Sokolova and Lapalme 2009).

The F1-score is defined as the harmonic mean of precision and recall and provides a more balanced performance indicator, particularly in fine-grained classification tasks where the cost of misclassification is high. Precision and recall are computed as follows (Sokolova and Lapalme 2009; Chicco and Jurman 2020):

$$Precision = \frac{TP}{TN + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Using these two measures, the F1-score is computed as follows:

$$F1 - score = 2.\frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

In the multi-class classification scenario, the F1-score was evaluated by averaging the class-wise F1 values computed for each category. This macro-averaging strategy makes the impact of class imbalance on model performance more explicit and provides a more sensitive reflection of misclassifications, particularly among visually similar classes (Sokolova and Lapalme 2009; Chicco and Jurman 2020).

In addition to classification performance, computational efficiency was also incorporated into the evaluation process (Sokolova and Lapalme 2009). However, the absence of additional efficiency indicators such as FLOPs, memory consumption, and throughput limits the completeness of the efficiency analysis. In this context, the total number of parameters and the inference time of each model were

measured (Chicco and Jurman 2020). While these metrics provide a practical estimation of computational cost, they do not fully capture all aspects of efficiency, such as floating-point operations (FLOPs), memory consumption, and throughput. The parameter count provides direct insight into the model's memory requirements and deployment cost, whereas inference time reflects its practical usability in real-time or latency-sensitive applications. Inference time is defined as the duration required for a single forward pass of the trained model (Chicco and Jurman 2020).

Through this multi-dimensional evaluation framework, the models were compared not only in terms of achieving the highest classification accuracy, but also with respect to the trade-off between accuracy, computational speed, and model complexity. Nevertheless, a more comprehensive efficiency analysis could be achieved by incorporating additional metrics such as FLOPs, memory usage, and frames per second (FPS), which are particularly relevant for deployment-oriented scenarios. When considered alongside the transfer learning strategies, architecture-aware learning rate adjustments, and unified optimization framework described in the previous sections, these evaluation metrics enable the observed performance differences to be directly attributed to architectural representation mechanisms. Consequently, the study provides a more comprehensive and practical assessment of which architectural approaches are most suitable for fine-grained image classification tasks under different application constraints. Future work should integrate these metrics to provide a more comprehensive and system-level understanding of computational efficiency across architectures.

Results

The experimental results obtained on the Oxford 102 Flowers dataset provide a multi-dimensional evaluation of fine-grained image classification performance across convolutional, hybrid, and transformer-based architectural paradigms. Since all models were evaluated under identical data splits, preprocessing procedures, and a unified training–optimization framework, the observed performance differences can be attributed directly to architectural representation capacity and computational characteristics rather than data-dependent confounding factors. This controlled setting enables a joint assessment of classification accuracy, model complexity, and inference latency.

Table 1 summarizes the classification performance and computational costs of all evaluated models. When accuracy, F1-score, parameter count, and inference time are jointly considered, the Vision Transformer–based ViT-L/16 achieves the highest classification performance, attaining an accuracy of 99.37% and an F1-score of 99.27%. Given the pronounced inter-class visual similarity and fine-grained nature of the Oxford 102 Flowers dataset, this result reflects a high level of discriminative capability within this specific benchmark setting, rather than indicating a generalizable superiority across different datasets.

However, Table 1 also shows that this performance is obtained at a substantial computational cost. ViT-L/16 contains 303.4 million parameters and incurs an inference latency of 27.4 ms per sample, making it the most computationally expensive and slowest model evaluated in this study. Consequently, ViT-L/16 represents an upper bound in terms of absolute accuracy, while simultaneously defining an extreme point in the accuracy–efficiency trade-off.

This behavior is consistently reflected in the test accuracy comparison shown in Fig. 2, where ViT-L/16 occupies the top position along the accuracy axis but is clearly associated with increased model complexity relative to other architectures.

Among hybrid models, ConvNeXt-V2-Tiny constitutes one of the most notable findings of the study. As reported in Table 1, ConvNeXt-V2-Tiny achieves an accuracy of 98.13% and an F1-score of 98.05%, approaching the performance of ViT-L/16 while requiring only 27.9 million parameters and

Table 1 Classification Performance and Computational Costs for All Evaluated Models

Model	Architecture Type	Accuracy (%)	F1-Score (%)	Parameters (M)	Inference Time (ms)
ViT-L/16	ViT	99.37	99.27	303.4	27.40
ConvNeXt-V2-Tiny	Hybrid	98.13	98.05	27.9	0.77
ViT-B/16	ViT	97.04	96.93	85.9	1.22
MaxViT-Tiny	Hybrid	93.27	93.03	30.5	1.26
DeiT-Base	ViT	92.63	92.79	85.9	1.14
Swin-Tiny	Hybrid	92.16	91.72	27.6	0.70
CoAtNet-0	Hybrid	91.87	91.61	26.7	1.05
VGG-16	CNN	88.83	88.23	134.7	0.97
EfficientNet-B0	CNN	88.05	87.75	4.1	0.31
ResNet-50	CNN	73.88	72.09	23.7	0.43

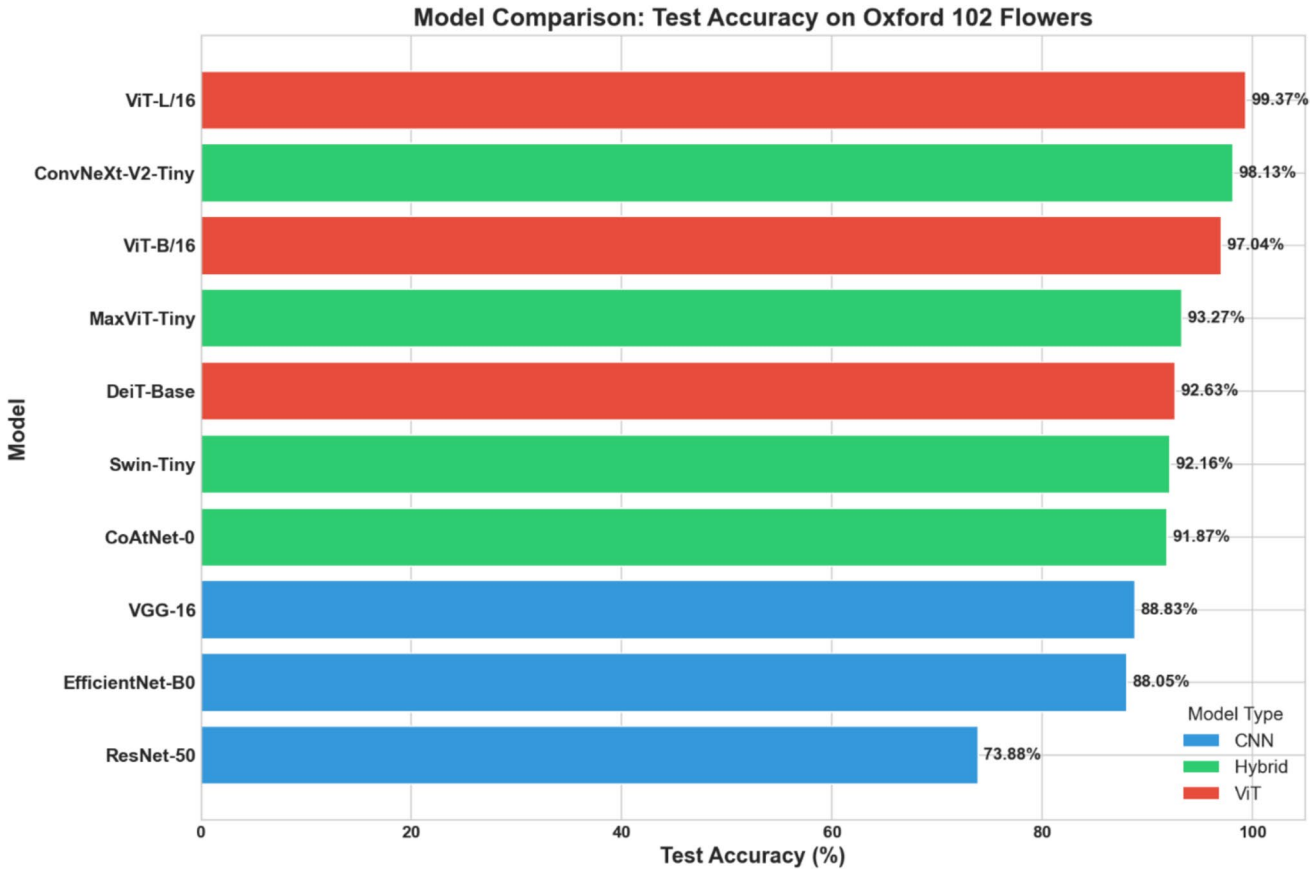

Fig. 2 Test accuracy comparison across all models (*Blue* CNN, *Green* Hybrid, *Red* ViT)

an inference time of 0.77 ms. Relative to ViT-L/16, this corresponds to a substantial reduction in parameter count and inference latency, with only a modest decrease in absolute accuracy.

Furthermore, ConvNeXt-V2-Tiny outperforms several Vision Transformer models, including ViT-B/16, in terms of accuracy (Table 1; Fig. 2). This observation demonstrates that hybrid architectures are not merely intermediate compromises between convolutional and transformer-based paradigms, but can deliver directly competitive performance when architectural design choices are appropriately optimized.

The remaining hybrid models—MaxViT-Tiny, Swin-Tiny, and CoAtNet-0—exhibit a compact range of parameter sizes (approximately 26–30 million) and achieve accuracy values between 91.87% and 93.27% (Table 2), indicating architectural consistency with respect to computational complexity.

CNN-based architectures display a wider variation in both performance and efficiency. VGG-16 and EfficientNet-B0 achieve comparable accuracy levels (88.83% and 88.05%, respectively), despite a pronounced difference in parameter count. While VGG-16 comprises 134.7 million parameters, EfficientNet-B0 attains similar accuracy with only 4.1 million parameters and an inference time of 0.31 ms (Table 1). This comparison highlights substantial differences in parameter efficiency within the CNN paradigm.

Table 2 Internal comparison of hybrid models

Model	Accuracy (%)	Parameters (M)	Inference Time (ms)
ConvNeXt-V2-Tiny	98.13	27.9	0.77
MaxViT-Tiny	93.27	30.5	1.26
Swin-Tiny	92.16	27.6	0.70
CoAtNet-0	91.87	26.7	1.05

In contrast, ResNet-50 exhibits the lowest classification performance among all evaluated models, achieving an accuracy of 73.88% and an F1-score of 72.09%. As illustrated in Fig. 1, ResNet-50 is clearly separated from both hybrid and transformer-based architectures in the accuracy distribution, indicating limited representational capacity for highly fine-grained classification tasks.

Average performance values across architectural paradigms are reported in Table 3. Vision Transformer models achieve the highest average accuracy (96.35%), followed by

Table 3 Average performance comparison across architectural paradigms

Architecture Type	Average Accuracy (%)	Average Parameters (M)	Average Inference Time (ms)
CNN	83.59	54.2	0.57
Hybrid	93.86	28.2	0.95
ViT	96.35	158.4	9.92

hybrid architectures (93.86%), while CNN-based models exhibit a substantially lower average accuracy (83.59%). These results are consistent with the observation that architectures capable of capturing broader contextual representations tend to achieve higher performance in fine-grained classification settings. Nevertheless, the average values alone do not capture the intra-group performance variability.

This relationship between accuracy and model complexity is further illustrated in Fig. 3, which jointly visualizes accuracy and parameter count across models.

Figure 4 presents the distributions of test accuracy, inference time, and parameter count across the three architectural paradigms. In terms of accuracy, Vision Transformer models exhibit the highest median performance but also show a wider spread, reflecting substantial variability across different configurations. In contrast, hybrid architectures display a more compact accuracy distribution, indicating greater performance consistency within this group.

From a computational perspective, the inference time and parameter distributions in Fig. 4 reveal a clear separation between paradigms. CNN-based models cluster tightly around low latency and small parameter counts, whereas Vision Transformer models demonstrate both higher median values and markedly larger variance, driven primarily by large-scale architectures such as ViT-L/16. Hybrid models occupy an intermediate yet compact region, combining relatively low inference times with moderate parameter counts and reduced variance.

The inference time comparison presented in Fig. 5 shows that EfficientNet-B0 and ResNet-50 are the fastest models, with inference latencies in the range of approximately 0.3–0.4 ms (Table 1). Most hybrid architectures achieve inference times close to or below 1 ms, with ConvNeXt-V2-Tiny offering a favorable high-accuracy–high-speed trade-off at an inference time of 0.77 ms. In contrast, ViT-L/16 exhibits a substantially higher inference time of 27.4 ms, clearly separating it from all other models in terms of computational efficiency.

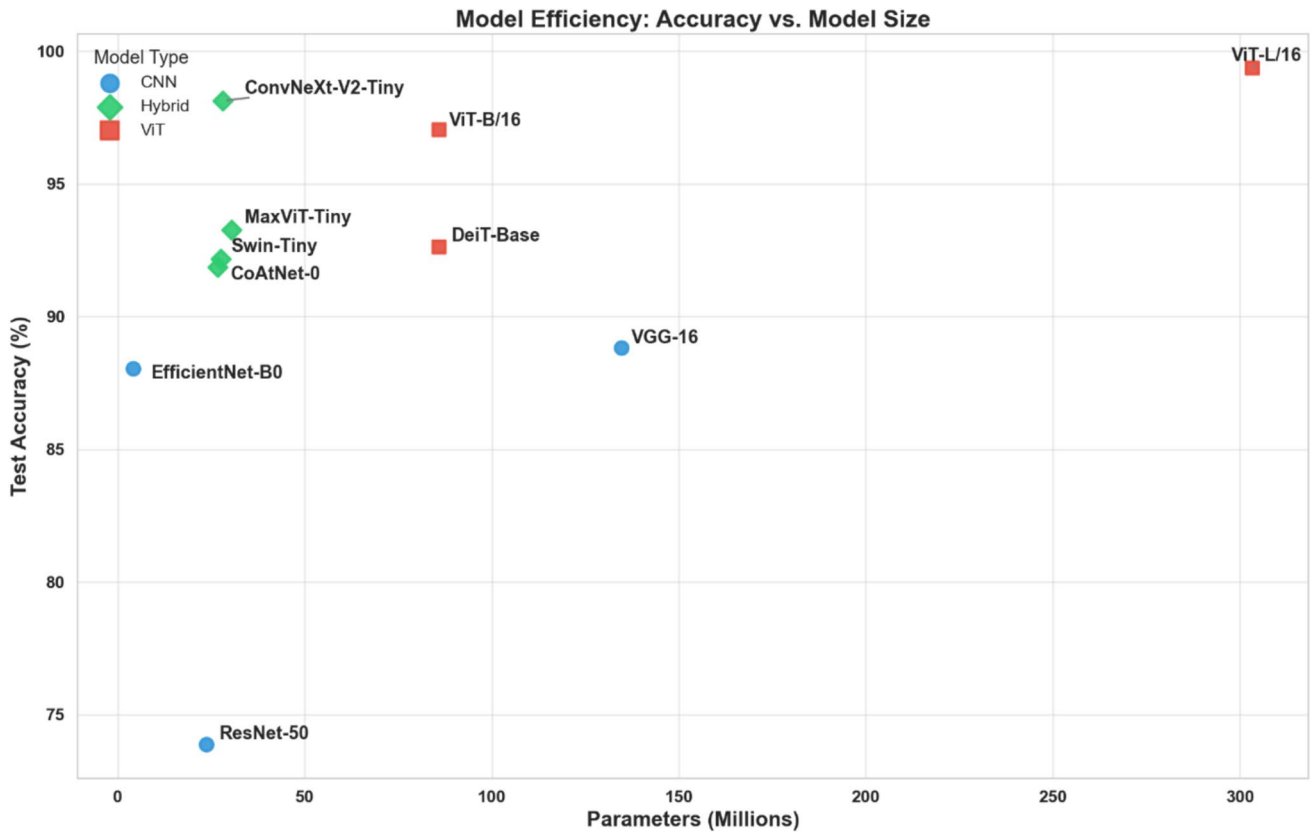

Fig. 3 Model efficiency–Accuracy vs. Parameters (size indicates relative model size)

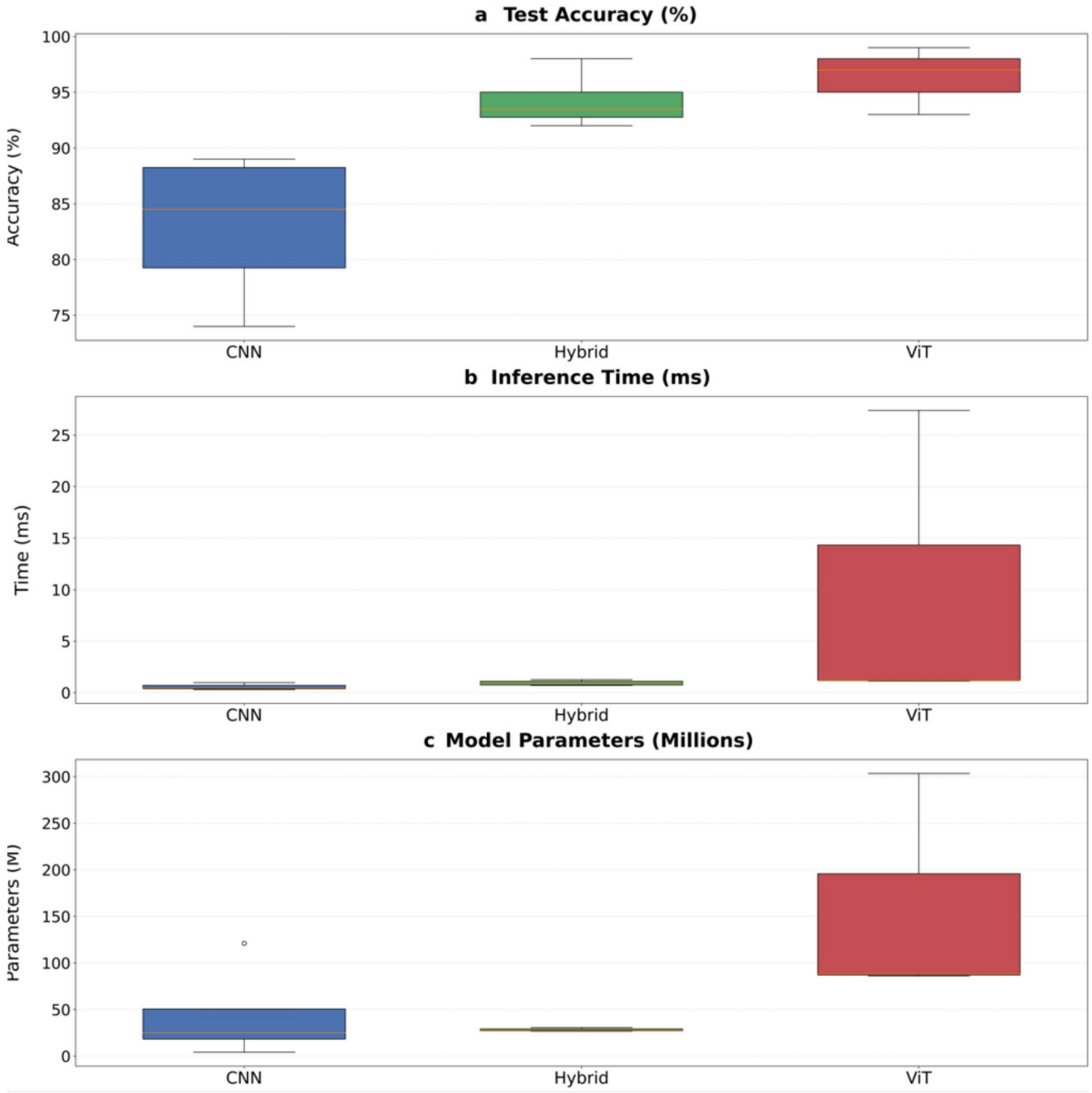

Fig. 4 Distribution of accuracy (**a**), inference time (**b**), and parameters by architecture type (**c**)

A focused comparison of the best-performing models is provided in Table 4, while an internal comparison of hybrid architectures is summarized in Table 2. When model accuracy is jointly analyzed with parameter count and inference time (Table 2; Fig. 3), hybrid architectures demonstrate a more balanced and predictable performance profile. Hybrid models cluster within a narrow parameter range while achieving high accuracy values, whereas Vision Transformer models span a substantially wider range of model sizes to attain peak accuracy.

Overall, these results indicate that, in fine-grained image classification tasks, the highest absolute accuracy does not necessarily correspond to the most balanced solution when computational efficiency is taken into account.

To further operationalize the performance–efficiency trade-off, Table 5 reports normalized ratios such as accuracy per 10 M parameters and macro-F1 per millisecond. While ViT-L/16 achieves the highest absolute accuracy, its normalized efficiency ratios are substantially lower than those of compact hybrid and CNN models. In contrast, ConvNeXt-V2-Tiny and

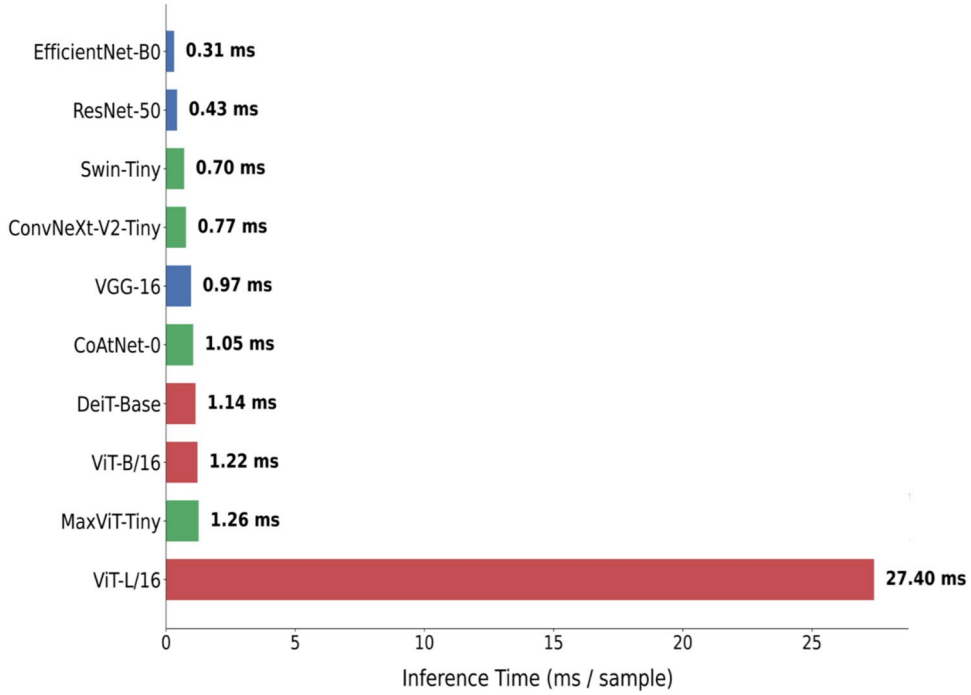

Fig. 5 Inference time comparison (lower is better)

Table 4 Accuracy–efficiency comparison of the best-performing models

Model	Accuracy (%)	Parameters (M)	Inference Time (ms)
ViT-L/16	99.37	303.4	27.40
ConvNeXt-V2-Tiny	98.13	27.9	0.77
EfficientNet-B0	88.05	4.1	0.31

ResNet-50 exhibit markedly higher efficiency-normalized scores, highlighting that deployment-oriented model selection may favor architectures with balanced or speed-optimized profiles rather than peak accuracy alone.

Discussion

The comparative analysis reveals distinct performance patterns among convolutional, hybrid, and transformer-based architectures in fine-grained image classification under the evaluated experimental setup. Compared to prior comparative studies that often aggregate results across heterogeneous datasets and experimental settings, the present work adopts a strictly controlled experimental design to minimize confounding factors. This approach enables a more direct interpretation of architectural behavior, although it necessarily limits the generalizability of the conclusions to broader and more diverse fine-grained datasets. The ViT-L/16 model achieved the highest overall performance with an accuracy of 99.37% and an F1-score of 99.27% within the evaluated experimental setting. A similar trend was observed for the F1-score, where ViT-L/16 (99.27%) exceeded ConvNeXt-V2-Tiny (98.05%) by 1.22 percentage points under identical evaluation conditions. However, the performance margin relative to hybrid architectures remains limited (≈1%), and should be interpreted cautiously given the near-saturated nature of the dataset.

However, this marginal accuracy gain is accompanied by a substantial increase in computational cost: ViT-L/16 requires 303.4 M parameters, approximately 10.9 × larger than ConvNeXt-V2-Tiny (27.9 M) and 2.25 × larger than

Table 5 Deployment-oriented efficiency indicators across architectures

Model	Acc (%)	Macro-F1 (%)	Params (M)	Latency (ms/img)	Acc per 10 M params	F1 per ms
ViT-L/16	**99.37**	**99.27**	**303.4**	**27.40**	**3.28**	**3.62**
ConvNeXt-V2-Tiny	98.13	98.05	27.9	0.77	35.17	127.34
VGG-16	88.83	88.23	134.7	0.97	6.59	90.96
ResNet-50	73.88	72.09	23.7	0.43	31.17	167.65

The numbers highlighted in bold are the highest modeled data points

VGG-16 (134.7 M), while incurring an inference latency of 27.4 ms—approximately 35 × slower than ConvNeXt-V2-Tiny (0.77 ms). From a deployment-oriented perspective, this raises an important practical question: whether an absolute improvement of approximately 1% in accuracy justifies an order-of-magnitude increase in model size and inference time. Such trade-offs between performance scaling and model complexity have been widely discussed in the context of scaling laws and diminishing returns in deep learning (Kaplan et al. 2020). In resource-constrained environments, such trade-offs may favor architectures that offer slightly lower peak accuracy but substantially improved efficiency. Efficiency-aware evaluation has increasingly been emphasized as a critical component of model assessment in applied machine learning systems (Maier-Hein et al. 2024). In saturation regimes, incremental gains in accuracy often require disproportionately larger model capacity, suggesting that marginal performance improvements may reflect scaling dynamics rather than fundamentally different representational mechanisms.

Within the hybrid paradigm, ConvNeXt-V2-Tiny consistently outperformed Swin-Tiny, MaxViT-Tiny, and CoAtNet-0 by 4.86–6.26 percentage points in accuracy, despite maintaining a comparable parameter range (≈26–30 M). This suggests that specific architectural refinements within hybrid frameworks may influence performance more strongly than parameter count alone. In contrast, CNN-based models exhibited lower accuracy values under the evaluated configuration: EfficientNet-B0 and VGG-16 remained below 89%, while ResNet-50 achieved 73.88%. These findings should be interpreted strictly within the context of the Oxford 102 Flowers dataset under transfer learning and standardized resolution, rather than as evidence of general architectural inadequacy across fine-grained tasks.

Collectively, the results indicate that while large-scale Vision Transformers achieve the highest absolute accuracy within this benchmark, hybrid architectures—particularly ConvNeXt-V2-Tiny—provide a more balanced accuracy–efficiency profile under the same experimental conditions. Importantly, Oxford 102 Flowers represents a relatively mature and near-saturated fine-grained benchmark under transfer learning (Nilsback and Zisserman 2008; Wei et al. 2021); therefore, performance differences below approximately 2% at very high accuracy levels should be interpreted with caution in terms of practical significance. This observation is consistent with recent studies emphasizing efficiency-aware evaluation and scaling behavior, where marginal gains in accuracy are often associated with substantial increases in computational cost (Rusia and Singh 2023; Alsalihi and Sztahó 2024; Raposo 2024; Ferreira and Couto 2025a, b, c; Ferreira et al. 2025b). Similarly, contemporary research on hybrid architectures suggests that combining convolutional inductive biases with attention mechanisms can achieve competitive performance with improved efficiency. In this context, the strong performance of ConvNeXt-V2-Tiny aligns with recent findings highlighting the effectiveness of hybrid design strategies under deployment constraints (Apap et al. 2024; Ferreira et al. 2024, 2025a; Ferreira and do Couto 2025b, 2025a).

At a representational level, the observed performance differences can be interpreted through the balance between local feature induction and global context modeling. Classical CNNs rely on convolutional inductive biases such as spatial locality and translation equivariance (He et al. 2016; Younesi et al. 2024). In fine-grained classification settings, where inter-class differences may depend on subtle global arrangements or distributed color–shape correlations (Wei et al. 2021), purely locality-driven representations may face limitations. Vision Transformers address this by employing global self-attention mechanisms (Vaswani et al. 2017; Dosovitskiy et al. 2021), enabling each patch to attend to all others. However, reduced inductive bias can increase data and computational demands (Khan et al. 2022). This contrast highlights a structural trade-off between inductive bias strength and representational flexibility.

Hybrid architectures integrate convolutional inductive biases with attention-based modeling (Chen et al. 2025). Models such as ConvNeXt-V2-Tiny preserve locality in early stages while incorporating broader contextual modeling mechanisms in deeper layers (Mmileng et al. 2025; Poyraz et al. 2025). This architectural combination appears to yield stable optimization and competitive performance under moderate computational cost within this dataset. Such designs may be interpreted as attempts to approximate global modeling capacity while retaining the regularization benefits of convolutional priors.

The present findings suggest that fine-grained image classification under controlled benchmark conditions does not strictly favor purely local or purely global representations. Rather, performance outcomes appear to depend on how effectively architectural inductive biases are aligned with the structural characteristics of the dataset. In this sense, architectural effectiveness should be interpreted as context-dependent rather than universally hierarchical.

Class-level and failure mode analysis

While aggregate accuracy and F1-score provide a global performance overview, fine-grained recognition literature emphasizes that class-level behavior and error distribution are critical for meaningful architectural comparison (Wah et al. 2011). Fine-grained benchmarks are characterized by subtle inter-class differences, where misclassification patterns often reveal representational limitations not visible in

overall accuracy metrics. However, these observations are based on quantitative indicators, and visual interpretability tools would be required to directly validate which image regions contribute to model decisions.

Consistent with prior observations in fine-grained evaluation studies (Wah et al. 2011), class-wise F1-scores and normalized confusion matrices were analyzed across representative architectures. The results indicate that performance disparities are not uniformly distributed across all categories but are concentrated in visually overlapping flower species exhibiting subtle variations in petal geometry, symmetry, and chromatic gradients.

CNN-based architectures display greater variance in per-class F1-scores, particularly in categories where discriminative information is globally distributed rather than locally textured. This observation aligns with the known inductive bias of convolutional models toward localized receptive fields and translation-equivariant representations.

In contrast, transformer-based architectures, grounded in global self-attention mechanisms (Vaswani et al. 2017), demonstrate improved recall in classes defined by holistic spatial configuration. However, qualitative inspection of misclassified samples reveals that transformer models occasionally overemphasize background context or peripheral regions, suggesting sensitivity to global contextual correlations beyond strictly discriminative object features.

Hybrid architectures exhibit comparatively lower per-class performance variance, indicating a more stable integration of convolutional locality and attention-driven global modeling. The normalized confusion matrices further reveal systematic confusion clusters among morphologically similar species, suggesting that certain error modes arise from intrinsic dataset ambiguity rather than purely architectural deficiencies.

These findings demonstrate that architectural differences manifest not only in aggregate performance but also in distinct error typologies, reinforcing the importance of class-level analysis in fine-grained evaluation.

Practical significance and cost–benefit analysis

Although ViT-L/16 achieves the highest absolute accuracy, the observed margin relative to ConvNeXt-V2-Tiny remains approximately 1.24%. Given the near-saturated nature of the Oxford 102 Flowers benchmark under transfer learning, such differences should be interpreted in terms of practical relevance rather than solely statistical magnitude.

Recent large-scale model scaling analyses have shown that performance improvements tend to exhibit diminishing returns as model size increases, particularly when approaching saturation regimes (Kaplan et al. 2020). In the present comparison, the additional 1.24% accuracy obtained by ViT-L/16 is accompanied by an increase in approximately 275 million parameters relative to ConvNeXt-V2-Tiny. This disproportionate growth in model size highlights a declining efficiency in parameter utilization at the upper end of the performance spectrum.

To characterize this trade-off in practical terms, we evaluate how much additional accuracy is gained relative to the substantial increase in model complexity. The results indicate that the incremental accuracy benefit per unit increase in parameters is considerably smaller than the gains observed when transitioning from conventional CNN architectures to hybrid models. This pattern is consistent with diminishing-return behavior reported in contemporary scaling studies (Kaplan et al. 2020).

From a deployment-oriented perspective, particularly in memory- or latency-constrained environments, an absolute improvement below 2% may not justify an order-of-magnitude increase in parameter count and inference cost. However, this interpretation remains context-dependent and may vary across application domains where peak predictive accuracy outweighs computational considerations. Consequently, architectural superiority should not be defined exclusively by peak benchmark accuracy, but rather by the balance between predictive performance and computational efficiency. In this context, peak accuracy ranking is effectively replaced by deployment-aware dominance analysis, where architectural preference is determined by efficiency-normalized trade-offs rather than absolute performance alone.

Transformer-based architectures may therefore be advantageous in research scenarios where computational constraints are minimal. In contrast, hybrid architectures offer a more balanced and deployment-friendly alternative for real-time or edge-based applications. This decision-oriented interpretation aligns with recent discussions emphasizing efficiency-aware evaluation in applied deep learning systems.

Limitations and transferability

Several limitations of this study should be explicitly acknowledged. First, all experiments were conducted exclusively on the Oxford 102 Flowers dataset (Nilsback and Zisserman 2008), which, although widely adopted in fine-grained recognition research, represents a moderately sized benchmark under transfer learning. While this dataset provides a controlled evaluation setting, its near-saturated nature under transfer learning leads to very high accuracy levels, where performance differences on the order of ~ 1% may not reflect substantial differences in practical discriminative capability. In such regimes, small accuracy improvements should be interpreted cautiously, as they may be influenced by dataset-specific characteristics rather than fundamental architectural advantages. Compared to more structurally complex fine-grained datasets such as

CUB-200 (Wah et al. 2011)or FGVC-Aircraft (Maji et al. 2013), Oxford 102 exhibits comparatively limited scale and relatively controlled object presentation. Consequently, the very high accuracy values observed, particularly for large transformer-based models, may partially reflect dataset-specific characteristics rather than universally generalizable architectural advantages.

In addition, although all models were trained under a unified experimental pipeline with identical data splits, standardized preprocessing, and comparable optimization schedules, such standardization reduces, but does not eliminate, potential sources of bias. In particular, architecture-specific hyperparameter choices, such as learning rate selection, may introduce implicit bias, as different models may benefit unequally from tailored optimization settings. Recent discussions on reproducibility and benchmarking in machine learning emphasize that architectural comparisons remain sensitive to pretraining scale, optimization dynamics, and hyperparameter configurations (Dodge et al. 2019; Recht et al. 2019). Therefore, the unified protocol adopted in this study should be interpreted as a mechanism to minimize experimental confounds rather than as a guarantee of absolute fairness across fundamentally different architectural families. Furthermore, the use of architecture-aware learning rate schedules introduces an implicit optimization asymmetry, as certain architectural families may exhibit different sensitivity profiles to hyperparameter configurations. Therefore, observed performance differences cannot be attributed exclusively to architectural inductive biases without considering optimization-related effects. Future work should include sensitivity analyses or unified hyperparameter experiments to assess how performance varies under consistent optimization settings across architectures.

Second, all models were pretrained on ImageNet-1 K or ImageNet-21 K (Russakovsky et al. 2015), and architecture-aware learning rate configurations were employed. While such pretraining strategies are standard practice and improve convergence stability, they inherently introduce optimization priors that may benefit certain architectural families more than others. In particular, transformer-based models have been shown to rely more heavily on large-scale pretraining to achieve competitive performance (Vaswani et al. 2017; Dosovitskiy et al. 2021). Therefore, the comparative results reported here should be interpreted as conditional on the adopted transfer learning regime rather than as architecture-intrinsic performance guarantees.

Third, experiments were conducted at a fixed input resolution of 224×224 pixels. Prior research has demonstrated that both convolutional and transformer-based models exhibit sensitivity to input resolution and scaling strategies (Touvron et al. 2021). It is therefore plausible that relative architectural advantages may shift under higher-resolution settings or alternative multi-scale training protocols.

Although the comparative results are reported under a controlled and fixed split evaluation, the present study reflects a single-run benchmarking design under the available experimental artifacts. While this design enables consistent and comparable evaluation across architectures, it does not account for stochastic variability arising from random initialization and training dynamics. Consequently, the absence of repeated runs with different random seeds limits the ability to perform formal statistical inference. Formal inferential statistics based on repeated-seed training (e.g., 5–10 seeds) and/or cross-validation would further strengthen claims by quantifying run-to-run variability and enabling significance testing across models. We therefore interpret performance gaps as practically meaningful when differences are large (e.g., CNN-to-transformer or hybrid gaps), while treating marginal improvements in the near-saturated regime (e.g., $\approx 1\%$ differences among top-performing models) primarily in terms of practical relevance and deployment constraints rather than statistical magnitude alone. In particular, small performance differences ($\approx 1\%$) observed in near-saturated settings should not be overinterpreted without statistical validation, as they may fall within expected run-to-run variability. This consideration is particularly relevant in near-saturated datasets, where high accuracy values may mask underlying class-level ambiguities and reduce the interpretability of small performance differences. Although formal repeated-seed statistical testing was not conducted within the present controlled single-run framework, the magnitude of inter-family performance gaps (e.g., CNN vs. Vision Transformer paradigms) substantially exceeds typical run-to-run variance reported in comparable transfer learning studies. Therefore, large architectural disparities are interpreted as practically robust patterns, whereas marginal differences in the near-saturated regime are treated cautiously. Future work should incorporate multiple experimental runs (e.g., ≥ 5 seeds), report mean performance with standard deviation, and apply statistical significance tests such as t-tests or non-parametric alternatives to provide a more rigorous assessment of performance differences.

Finally, the conclusions are restricted to image-level classification under controlled experimental conditions. Transferability to more complex fine-grained benchmarks—such as CUB-200 (Wah et al. 2011), FGVC-Aircraft (Maji et al. 2013), or large-scale ecological datasets—requires further empirical validation. Additionally, the absence of class-level evaluation metrics limits the ability to fully analyze misclassification patterns and category-specific performance differences. Future work should extend the present comparative framework across multiple datasets, varying resolution regimes, and alternative pretraining strategies in order to rigorously assess robustness and generalizability. Future work should incorporate confusion matrix analysis, per-class precision and recall,

and balanced accuracy to enable a more comprehensive and interpretable evaluation of fine-grained classification performance.

Model-wise qualitative error inspection and attribution-based interpretability analyses (e.g., Grad-CAM visualizations) constitute important complementary dimensions for examining failure modes and representational focus in deep neural networks (Selvaraju et al. 2017). However, recent work has demonstrated that saliency and attribution methods can be sensitive to implementation details and model randomization, raising concerns regarding attribution artifacts and interpretability reliability (Adebayo et al. 2018; Samek et al. 2021). Meaningful cross-architecture comparison of attribution maps therefore requires strictly harmonized preprocessing pipelines, checkpoint alignment, and identical inference codepaths to prevent spurious differences that may reflect experimental variance rather than intrinsic representational behavior. In the absence of such fully synchronized artifacts within the present revision scope, systematic cross-model misclassification panels and standardized XAI-based attention comparisons are identified as methodologically rigorous extensions of this work. In particular, visual explanation techniques such as Grad-CAM and attention map analysis could provide valuable insights into discriminative regions used by different architectures, as well as reveal systematic failure cases in visually similar classes. Such qualitative analyses would enhance the interpretability of model decisions and support a more comprehensive evaluation beyond quantitative metrics.

Furthermore, although the current benchmarking framework ensures controlled single-run comparability, recent analyses in machine learning reproducibility have shown that performance estimates may vary substantially across random seeds, optimization dynamics, and minor perturbations (Dodge et al. 2019; Recht et al. 2019). Incorporating repeated-seed training and cross-validation designs would therefore strengthen inferential robustness by quantifying variability and enabling formal statistical comparisons across architectural families. Such variance-aware evaluation constitutes a natural next step toward enhancing the statistical reliability and generalizability of comparative deep learning studies. Furthermore, while the controlled single-run experimental design ensures comparability, future studies incorporating repeated trials, statistical significance testing, and cross-validation protocols would further strengthen the robustness and generalizability of the findings.

Positioning against recent comparative and deployment-oriented studies

Recent comparative and survey-oriented studies (2023–2025) have extensively examined the evolution of convolutional, transformer-based, and hybrid architectures across multiple visual domains (Khan et al. 2022; Takahashi et al. 2024; Hassija et al. 2025). These works primarily provide broad taxonomic overviews, reporting aggregated performance trends across heterogeneous datasets. In contrast, the present study does not aim to synthesize architectural evolution at a survey level, but instead isolates architectural trade-offs under a strictly controlled experimental protocol with fixed splits, unified preprocessing, and architecture-aware optimization. This controlled isolation allows performance differences to be interpreted with reduced experimental confounding, thereby shifting the emphasis from descriptive benchmarking toward deployment-oriented efficiency analysis. Consistent with recent studies, our results show that marginal accuracy gains often require disproportionate increases in model complexity. In line with hybrid architecture literature, ConvNeXt-V2-Tiny achieves competitive performance with significantly lower computational cost. Unlike prior studies, our controlled setup enables a direct comparison of architectural paradigms without confounding factors(Apap et al. 2024; Ferreira et al. 2024, 2025a; Ferreira and do Couto 2025b, 2025a).

Several recent domain-specific hybrid studies (Chen et al. 2025; Lilhore et al. 2025; Wu et al. 2025), demonstrate that combining convolutional inductive biases with attention mechanisms improves robustness in medical or domain-constrained imaging tasks. However, these works typically introduce task-specific architectural modifications or multi-branch fusion mechanisms. In contrast, the present study deliberately evaluates off-the-shelf representative architectures without architectural customization. This distinction is important: rather than proposing incremental hybrid refinements, our contribution lies in quantifying how existing architectural families behave under identical fine-grained benchmark constraints.

Recent works emphasizing efficiency and adaptability (Maier-Hein et al. 2024; Yang 2025) argue that peak benchmark accuracy alone is insufficient for assessing model utility, particularly in deployment-constrained environments. The present findings empirically support this perspective by explicitly quantifying diminishing accuracy gains relative to parameter growth. Unlike general efficiency discussions, however, our analysis provides concrete trade-off ratios (e.g., accuracy gain per parameter increase) within a near-saturated fine-grained benchmark regime, thereby operationalizing efficiency-aware evaluation rather than merely advocating it conceptually.

Explainability-oriented studies (Aburass et al. 2025; Kumru et al. 2025a, 2025b) increasingly stress the importance of model interpretability and failure analysis. While the present work does not integrate XAI modules directly into the training pipeline, it extends beyond aggregate accuracy reporting by incorporating class-level F1 analysis and structured failure mode discussion. This analytical extension

differentiates the study from purely performance-driven comparisons and aligns it with emerging reliability-centered evaluation paradigms.

Finally, while large-scale surveys and cross-domain transformer analyses (Hassija et al. 2025; Mehdipour et al. 2025) often conclude that transformer-based architectures dominate in high-capacity regimes, such conclusions are typically derived from multi-dataset aggregation. The present study demonstrates that within a near-saturated fine-grained dataset under transfer learning, marginal accuracy improvements may reflect scaling dynamics rather than fundamental representational breakthroughs. In this respect, the work refines rather than contradicts existing literature by contextualizing architectural superiority within deployment and saturation constraints.

Collectively, these contrasts clarify that the contribution of the present study is neither architectural innovation nor large-scale meta-analysis. Instead, it provides a controlled, efficiency-aware, and deployment-oriented interpretation of architectural trade-offs within a fine-grained benchmark regime, thereby complementing rather than duplicating recent comparative literature.

Conclusions

This study provides a systematic and architecture-centric evaluation of convolutional neural networks, Vision Transformers, and hybrid deep learning models for fine-grained image classification using the Oxford 102 Flowers dataset. The experimental results demonstrate that large-scale Vision Transformers establish the upper bound of classification performance, with ViT-L/16 achieving 99.37% accuracy and 99.27% F1-score, suggesting the effectiveness of global self-attention within the evaluated dataset and experimental configuration. However, this marginal performance gain comes at a substantial computational cost, including a $>10\times$ increase in parameter count and a $>35\times$ increase in inference latency compared to efficient hybrid models. In contrast, hybrid architectures, most notably ConvNeXt-V2-Tiny, achieve a near-optimal balance between accuracy and efficiency, reaching 98.13% accuracy with an order-of-magnitude reduction in model size and inference time. Conventional CNN-based models achieved lower performance under the evaluated transfer learning configuration on the Oxford 102 Flowers dataset, suggesting potential limitations in capturing extremely subtle inter-class variations within this specific benchmark setting rather than indicating universal architectural inadequacy. Collectively, these findings demonstrate that the highest absolute accuracy does not necessarily correspond to the most practically viable solution, and that within the evaluated transfer learning configuration on Oxford 102 Flowers.

Beyond the reported results, the findings of this work suggest several important directions for future research. First, the Oxford 102 Flowers dataset should be re-evaluated using alternative learning paradigms, including lightweight transformers, state-space models, and ensemble-based strategies, to further explore the limits of accuracy–efficiency trade-offs under controlled conditions. Second, the integration of explainable artificial intelligence (XAI) techniques, such as attention visualization, Grad-CAM, or concept-based attribution methods, should be systematically investigated to elucidate how different architectures localize and utilize class-discriminative visual cues in fine-grained settings. Such analyses are essential for model transparency, reliability, and informed architectural selection. Finally, the architectural insights gained from this study are not limited to botanical imagery; the evaluated models and methodological framework can be directly extended to other biological data modalities, including macro- and micro-scale organism images, histopathological slides, cellular microscopy, pollen and spore morphology, and multimodal biological datasets. Applying these architectures across diverse biological domains will contribute to more generalizable, interpretable, and biologically meaningful artificial intelligence systems.

Author contributions Conceptualization contributed by D.M.Ö., H.S., F.E., and Ö.D.; methodology contributed by F.E., K.A., Ö. D., M.S.G. and I.A.; software contributed by F.E., D.M.Ö., H.S. and M.S.G.; validation contributed by M.S.G., H.S and F.E.; formal analysis contributed by M.S.G, Ö.D, K.A., and I.A.; investigation contributed by F.E. and M.S.G.; resources contributed by D.M.Ö, and H.S.; writing—original draft preparation contributed by I.A., A.K., M.S.G, M.S., E.C.E. and F.E.; writing—review & editing contributed by K.A., M.S.G., M.S., E.C.E. and F.E.; visualization contributed by F.E., M.S.G. and K.A.; supervision contributed by M.S.G., K.A., I.A. and F.E. All authors have read and agreed to the published version of the manuscript.

Data availability The dataset used in this study is the Oxford 102 Flowers dataset, which is publicly available and can be accessed from the original source described by Nilsback and Zisserman (2008). All experiments were conducted using the official training, validation, and test splits provided with the dataset. No new data were generated during the current study. Any additional information required to reproduce the results is available from the corresponding author upon reasonable request.

Declarations

Conflicts of interest The authors declare no conflict of interest.

References

Aamir M, Rahman Z, Choudhry N, Bhutto JA, Abro WA, Zhu Z (2025) From CNNs to Transformers: a review of evolving deep learning architectures for brain tumor classification. IEEE Access 13:184918–184936. https://doi.org/10.1109/access.2025.3625607

Aburass S, Dorgham O, Al Shaqsi J, Abu Rumman M, Al-Kadi O (2025) Vision Transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases. J Imaging Inform Med 38:3928–3971. https://doi.org/10.1007/s10278-025-01481-y

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Curran Associates Inc., Red Hook, pp 9525–9536. https://doi.org/10.48550/arXiv.1810.03292

Alomar K, Aysel HI, Cai X (2025) CNNs, RNNs and transformers in human action recognition: a survey and a hybrid model. Artif Intell Rev 58:1–44. https://doi.org/10.48550/arXiv.2407.06162

Alsalihi MH, Sztahó D (2024) Effect of identical twins on deep speaker embeddings based forensic voice comparison. Int J Speech Technol 27:341–351. https://doi.org/10.1007/s10772-024-10108-6

Apap A, Bhole A, Fernández-Robles L, Castejón-Limas M, Azzopardi G (2024) Explainable multi-layer COSFIRE filters robust to corruptions and boundary attack with application to retina and palmprint biometrics. Neural Comput Appl 36:19231–19245. https://doi.org/10.1007/s00521-024-10164-8

Castellano G, Vessio G (2021) Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. Neural Comput Appl 33:12263–12282. https://doi.org/10.1007/s00521-021-05893-z

Chen J, Liang Z, Lu X (2025) A dual attention and cross layer fusion network with a hybrid CNN and transformer architecture for medical image segmentation. Sci Rep 15:35707. https://doi.org/10.1038/s41598-025-19563-w

Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:6. https://doi.org/10.1186/s12864-019-6413-7

Dai Z, Liu H, Le Q V, Tan M (2021) CoAtNet: Marrying convolution and attention for all data sizes. In: Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21). Curran Associates Inc., Red Hook, pp 3965–3977. https://doi.org/10.48550/arXiv.2106.04803

Dodge J, Gururangan S, Card D, Schwartz R, Smith NA (2019) Show your work: Improved reporting of experimental results. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 2185–2194. https://doi.org/10.18653/v1/D19-1224

Dosovitskiy A, Beyer L, Kolesnikov A (2021) An image is worth 16×16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, pp 1–21. https://doi.org/10.48550/arXiv.2010.11929

Ferreira FRT, Couto LMd (2025a) Development of a computational deep learning model for detecting people in aerial images and videos degraded by compression artifacts. Earth Sci Inform 18:408. https://doi.org/10.1007/s12145-025-01916-8

Ferreira FRT, do Couto LM (2025b) Explainable machine learning framework integrating PPPM for automated cytological triage and personalized breast cancer risk stratification. Netw Model Anal Health Inform Bioinform 14:167. https://doi.org/10.1007/s13721-025-00669-0

Ferreira FRT, do Couto LM (2025c) Using deep learning on microscopic images for white blood cell detection and segmentation to assist in leukemia diagnosis. J Supercomput 81:410. https://doi.org/10.1007/s11227-024-06903-2

Ferreira FRT, do Couto LM, de Melo Baptista Domingues G (2024) Exploring the potential of YOLOv8 in hybrid models for facial mask identification in diverse environments. Neural Comput Appl 36:22037–22052. https://doi.org/10.1007/s00521-024-10351-7

Ferreira FRT, do Couto LM, de Melo Baptista Domingues G (2025a) Comparing the efficiency of YOLO-M for face recognition in images and videos degraded by compression artifacts. Evol Syst 16:70. https://doi.org/10.1007/s12530-025-09699-5

Ferreira FRT, do Couto LM, de Melo Baptista Domingues G, Saporetti CM (2025b) Development of a framework using deep learning for the identification and classification of engagement levels in distance learning students. Soc Netw Anal Min 15:37. https://doi.org/10.1007/s13278-025-01408-z

Hassija V, Palanisamy B, Chatterjee A, Mandal A, Chakraborty D, Pandey A, Kumar D (2025) Transformers for Vision: A Survey on Innovative Methods for Computer Vision. IEEE Access 13:95496–95523. https://doi.org/10.1109/ACCESS.2025.3571735

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling laws for neural language models. https://doi.org/10.48550/arXiv.2001.08361

Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2022) Transformers in vision: a survey. ACM Comput Surv 54:1–41. https://doi.org/10.1145/3505244

Korkmaz AF, Ekinci F, Altaş Ş, Kumru E, Güzel MS, Akata I (2025) A deep learning and explainable AI-based approach for the classification of Discomycetes species. Biology 14:719. https://doi.org/10.3390/biology14060719

Kumru E, Ekinci F, Aydoğan A (2025a) Explainable convolutional neural network architectures for high-performance taxonomic classification of gasteroid macrofungi. Sci Rep 15:40196. https://doi.org/10.1038/s41598-025-23974-0

Kumru E, Korkmaz AF, Ekinci F, Aydoğan A, Güzel MS, Akata I (2025b) Deep ensemble learning and explainable AI for multiclass classification of Earthstar fungal species. Biology 14:1313. https://doi.org/10.3390/biology14101313

Li Z, Hu J, Wu K, Miao J, Zhao Z, Wu J (2024) Local feature acquisition and global context understanding network for very high-resolution land cover classification. Sci Rep 14:12597. https://doi.org/10.1038/s41598-024-63363-7

Lilhore UK, Sharma YK, Shukla BK, Vadlamudi MN, Simaiya S, Alroobaea R, Baqasah AM (2025) Hybrid convolutional neural network and bi-LSTM model with EfficientNet-B0 for high-accuracy breast cancer detection and classification. Sci Rep 15:12082. https://doi.org/10.1038/s41598-025-95311-4

Liu Z, Lin Y, Cao Y (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, pp 9992-10002. https://doi.org/10.48550/arXiv.2103.14030

Loshchilov I, Hutter F, (2019). Decoupled weight decay regularization. https://doi.org/10.48550/arXiv.2103.14030. https://doi.org/10.48550/arXiv.1711.05101

Luo J, Qu H, Zhao Y, Zhang J, Yang Y (2025) Deep learning for 3D fashion design: A survey from a sewing pattern-driven perspective. IEEE Transact Circuits Syst Video Technol 36:4734–4756. https://doi.org/10.1109/TCSVT.2025.3637565

Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, Jäger PF (2024) Metrics reloaded: recommendations for image analysis validation. Nature Meth 21:195–212. https://doi.org/10.1038/s41592-023-02151-z

Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A (2013) Fine-grained visual classification of aircraft. https://doi.org/10.48550/arXiv.1306.5151. https://doi.org/10.48550/arXiv.1306.5151

Manakitsa N, Maraslidis GS, Moysis L, Fragulis GF (2024) A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine

and robotic vision. Technologies 12:15. https://doi.org/10.3390/technologies12020015
Mehdipour S, Mirroshandel SA, Tabatabaei SA (2025) Vision transformers in precision agriculture: A comprehensive survey. Intell Syst Appl 29:200617. https://doi.org/10.1016/j.iswa.2025.200617.
Mmileng OP, Whata A, Olusanya M, Mhlongo S (2025) Application of ConvNeXt with transfer learning and data augmentation for malaria parasite detection in resource-limited settings using microscopic images. PLoS ONE 20:e0313734. https://doi.org/10.1371/journal.pone.0313734
Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP '08). IEEE Computer Society, Washington DC, pp 722–729. https://doi.org/10.1109/ICVGIP.2008.47
Poyraz M, Poyraz AK, Dogan Y, Gunes S, Mir HS, Paul JK, Acharya R (2025) BrainNeXt: novel lightweight CNN model for the automated detection of brain disorders using MRI images. Cogn Neurodyn 19:1–17. https://doi.org/10.1007/s11571-025-10235-z
Raposo VL (2024) When facial recognition does not 'recognise': erroneous identifications and resulting liabilities. AI Soc 39:1857–1869. https://doi.org/10.1007/s00146-023-01634-z
Recht B, Roelofs R, Schmidt L, Shankar V (2019) Do imagenet classifiers generalize to imagenet? In: Proceedings of the 36th International Conference on Machine Learning. PMLR 97, Long Beach, pp 5380–5400. https://doi.org/10.48550/arXiv.1902.10811
Rusia MK, Singh DK (2023) A comprehensive survey on techniques to handle face identity threats: challenges and opportunities. Multim Tools Appl 82:1669–1748. https://doi.org/10.1007/s11042-022-13248-6
Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115:211–252. https://doi.org/10.1007/s11263-015-0816-y
Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller K-R (2021) Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE 109:247–278. https://doi.org/10.1109/JPROC.2021.3060483
Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, pp 618–626. https://doi.org/10.1109/ICCV.2017.74
Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations 2015. https://doi.org/10.48550/arXiv.1409.1556
Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Managem 45:427–437. https://doi.org/10.1016/j.ipm.2009.03.002
Struniawski K, Machlanska A, Marasek-Ciolakowska A, Konopka A (2026) Automated pollen classification via subinstance recognition: a comprehensive comparison of classical and deep learning architectures. Appl Sci 16:720. https://doi.org/10.3390/app16020720
Takahashi S, Sakaguchi Y, Kouno N, Takasawa K, Ishizu K, Akagi Y, Hamamoto R (2024) Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. J Med Syst 48:84. https://doi.org/10.1007/s10916-024-02105-8
Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, PMLR 97, pp 6105–6114. https://doi.org/10.48550/arXiv.1905.11946
Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning, PMLR 139, pp 10347–10357. https://doi.org/10.48550/arXiv.2012.12877
Tu Z, Talebi H, Zhang H (2022) MaxViT: Multi-axis vision transformer. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds) Computer Vision – ECCV 2022. Lecture Notes in Computer Science, vol. 13684. Springer, Cham, pp 459–479. https://doi.org/10.1007/978-3-031-20053-3_27
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neu Inf Process Syst 30. https://doi.org/10.48550/arXiv.1706.03762
Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds-200–2011 dataset. Technical Report CNS-TR-2011–001, California Institute of Technology
Wei XS, Song YZ, Mac Aodha O, Wu J, Peng Y, Tang J, Belongie S (2021) Fine-grained image analysis with deep learning: a survey. IEEE Trans Pattern Anal Mach Intell 44:8927–8948. https://doi.org/10.1109/tpami.2021.3126648
Woo S, Debnath S, Hu R (2023) ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, pp 16133–16142. https://doi.org/10.48550/arXiv.2301.00808
Wu B, Shi J, Wu Y, Tang N, Luo Y (2025) TransXSSM: A Hybrid Transformer State Space Model with Unified Rotary Position Embedding. https://doi.org/10.48550/arXiv.2506.09507.
Yang M (2025) Efficiency and adaptability of deep learning architectures: A performance comparison of CNN, transformer, and hybrid models. In: 25th International Conference on Computer Vision, Application and Algorithm (CVAA) 2025. IEEE, Ningbo, pp 1–4. https://doi.org/10.1109/CVAA66438.2025.11193320
Younesi A, Ansari M, Fazli M, Ejlali A, Shafique M, Henkel J (2024) A comprehensive survey of convolutions in deep learning: applications, challenges, and future trends. IEEE Access 12:41180–41218. https://doi.org/10.48550/arXiv.2402.15490
Zhang Q, Xu Y, Zhang J, Tao D (2023) Vitaev2: vision transformer advanced by exploring inductive bias for image recognition and beyond. Int J Comput Vis 131:1141–1162. https://doi.org/10.48550/arXiv.2202.10108
Zhang W, Ye N, Zhang L, Liu X, Wu H, Song A (2025) Machar: a frequency-aware mamba-convolution hybrid architecture for sensor-based human activity recognition. IEEE Int Thing J 13:7513–7526. https://doi.org/10.1109/JIOT.2025.3639422

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Demir Mesut Öztaş[1] · Hakan Salehi[1] · Özge Demir[2] · Koray Acici[3] · Mehmet Serdar Güzel[4] · Emre Cem Eraslan[5] · Mustafa Sevindik[5] · Ilgaz Akata[6] · Fatih Ekinci[7]

✉ Mustafa Sevindik
sevindik27@gmail.com

✉ Fatih Ekinci
fatihekinci@ankara.edu.tr

[1] Department of Artificial Intelligence, Ankara University, Ankara, Türkiye

[2] Vocational School, Big Data Analytics Program, Beykoz University, Istanbul, Türkiye

[3] Department of Artificial Intelligence and Data Engineering, Faculty of Engineering, Ankara University, 06830 Ankara, Türkiye

[4] Department of Computer Engineering, Faculty of Engineering, Ankara University, 06100 Ankara, Türkiye

[5] Department of Biology, Faculty of Engineering and Natural Sciences, Osmaniye Korkut Ata University, Osmaniye, Türkiye

[6] Department of Biology, Faculty of Science, Ankara University, 06100 Ankara, Türkiye

[7] Institute of Artificial Intelligence, Ankara University, 06100 Ankara, Türkiye