



A Comparative Study of Modern CNN and Transformer Architectures for Aerial Scene Classification Using Aerial Image Dataset

S Jayanthi^{1*}, M. A. Josephine Sathya², Nayani Sateesh³, Muthuvel Laxmikanthan⁴, Guguloth Ravi⁵,
V. Manojkumar⁶

¹ Department of Artificial Intelligence and Data Science, Faculty of Science and Technology (IcfaiTech), The ICFAI Foundation for Higher Education, Hyderabad 501203, India

² Department of Computational Studies, School of Computational and Physical Sciences, Kristu Jayanthi University, Bangalore 560077, India

³ Department of Computer Science and Engineering, CVR College of Engineering, Hyderabad 501510, India

⁴ Department of Computer Science, Moodlakatte Institute of Technology & Management, Moodalkatte 576217, India

⁵ Department of Data Science, TKR College of Engineering and Technology, Hyderabad 500097, India

⁶ Department of Cyber Security, School of Computing, SRM institute of Science & Technology Tiruchirappalli, Trichy 621105, India

Corresponding Author Email: drsjayanthicse@gmail.com

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ijssse.160812>

ABSTRACT

Received: 26 June 2026

Revised: 13 August 2026

Accepted: 21 August 2026

Available online: 31 August 2026

Keywords:

aerial scene classification, Aerial Image Dataset, remote sensing, vision transformers, deep convolutional neural networks, Swin Transformer, statistical benchmarking, land use land classification

Aerial scene classification is a foundational task in remote sensing image analysis. However, rigorous comparisons of modern convolutional neural networks (CNNs) and vision transformers remain limited in the literature. This study presents a systematic benchmark of six deep learning architectures, Vision Transformer (ViT-B/16), Swin Transformer (Swin-T), ConvNeXt-Base, MobileNetV3-L, DenseNet-121, and EfficientNet-B4 on the Aerial Image Dataset (AID). All models were fine-tuned using a unified protocol with ImageNet-pretrained weights and a stratified 70% training, 10% validation, and 20% held-out testing split, with five independent runs, each using a different random seed (42-46). Performance was evaluated using multiple classification metrics, bootstrap confidence intervals, pairwise statistical comparisons (including two one-sided tests (TOST) equivalence analysis), computational profiling, and representation-level analysis. Swin-T achieved the highest mean test accuracy ($97.88\% \pm 0.16\%$) and weighted F1 ($97.88\% \pm 0.15\%$). All models achieved high one-vs-rest multiclass ROC-AUC values ranging from 0.9985 to 0.9996. However, run-to-run variability differed by architecture, with ViT-B/16 having the highest standard deviation in accuracy at 1.31 percentage points. MobileNetV3-L achieved a competitive $97.51\% \pm 0.19\%$ accuracy using merely 4.24 million parameters and 0.23 GFLOPs. In contrast, ViT-B/16 achieved $96.38\% \pm 1.31\%$ accuracy with 85.82 million parameters and 11.29 GFLOPs. Uniform Manifold Approximation and Projection (UMAP) analysis revealed differences in learned feature organization across architectures, with ViT-B/16 exhibiting comparatively greater inter-class overlap. Our findings provide practical, statistically grounded guidance for architecture selection on RGB aerial scene classification benchmarks and offer reproducible baselines for future comparative studies.

1. INTRODUCTION

Automated scene understanding is indispensable for transforming massive Earth observation streams into actionable geospatial intelligence. Within this domain, aerial scene classification serves as a core mechanism for applications ranging from urban growth monitoring to disaster response [1-3]. In remote sensing image analysis, aerial scene classification is a foundational task that assigns semantic labels to discrete image patches based on their dominant spatial and contextual characteristics [4]. Unlike conventional natural image classification, aerial scene classification presents several unique challenges. Remote sensing scenes

frequently exhibit substantial intra-class variability arising from seasonal changes, illumination differences, imaging conditions, scale variations, and diverse spatial arrangements of constituent objects. Simultaneously, many scene categories exhibit high inter-class similarity because they share common land-cover elements and structural patterns [5, 6]. Furthermore, extracting reliable features from aerial images is difficult because objects appear at unpredictable angles, span multiple scales, and are embedded in complex environments. Thus, designing deep learning architectures capable of learning distinct and adaptable scene features remains a key focus of current research.

Early approaches used the manual feature extraction

techniques, such as Scale-Invariant Feature Transform, Histogram of Oriented Gradients, Local Binary Patterns, and Bag-of-Visual-Words representations [7-9]. Deep convolutional neural networks (CNNs) transformed this field by enabling models to automatically learn features directly from raw images. Following the success of transfer learning in remote sensing, architectures such as ResNet, DenseNet, EfficientNet, and MobileNetV3 have significantly improved both accuracy and computational efficiency [5, 7, 10, 11]. Consequently, CNN-based models became the dominant paradigm for aerial scene classification and related remote sensing applications.

More recently, transformer-based architectures have introduced a fundamentally different approach to visual representation learning. The ViT represents an image as a sequence of patch tokens and applies self-attention to model relationships among them [12]. The Swin Transformer (Swin-T) enhances this framework by adopting hierarchical feature representations and localized attention windows, which optimize computational scalability while preserving contextual modeling power [13-16]. ConvNeXt demonstrated that modernized CNN architectures can achieve performance comparable to transformers while retaining convolutional inductive biases [17]. These developments have increased interest in cross-paradigm comparisons, yet the relative merits of transformer and CNN architectures for aerial scene classification remain insufficiently characterized.

Despite these significant advances, several limitations remain in existing benchmarking literature. First, many comparative studies apply inconsistent preprocessing, augmentation, optimization, and data-splitting strategies across models. These differences make it difficult to determine whether performance variations are due to the architecture itself or to the training procedure [18, 19]. Second, model comparisons are often based primarily on classification accuracy, while statistical significance testing and uncertainty analysis are rarely performed [20, 21]. Consequently, relatively small numerical improvements are frequently interpreted as meaningful without formal evidence regarding their reliability. Third, deployment-oriented characteristics are often overlooked despite their practical importance in resource-constrained remote sensing environments. Fourth, relatively little attention has been devoted to representation-level analysis.

Finally, the sensitivity of performance rankings to stochastic training and model initialization is not always explicitly evaluated, leaving uncertainty about the reproducibility of observed differences across independent runs.

These considerations motivate a controlled comparative evaluation of modern CNN and transformer architectures for aerial scene classification. In particular, the relationship among predictive performance, run-to-run variability, computational efficiency, and learned feature organization requires a joint examination rather than assessment based on accuracy alone. This study therefore benchmarks six architectures on the Aerial Image Dataset (AID) [22], a widely used benchmark for aerial scene classification.

This work addresses previously identified methodological gaps through five specific design choices: (i) implementing a three-way stratified data partition with a held-out test set to reduce evaluation bias. (ii) Conducting five independent training runs for each architecture using distinct random seeds to quantify run-to-run variability. (iii) Performing formal

pairwise significance testing utilizing McNemar's test, the Wilcoxon signed-rank test, and two one-sided tests (TOST) for equivalence testing. (iv) Engaging in deployment-oriented computational profiling that encompasses parameter count, GFLOPs, inference latency, peak VRAM consumption, and throughput across various batch sizes. (v) quantitative feature-space analysis via Uniform Manifold Approximation and Projection (UMAP) with cluster-separation metrics to support classification-level findings.

The primary contributions of this work are as follows:

- Development of a unified benchmark framework for evaluating six representative modern CNN and transformer architectures.

- Integration of statistically rigorous evaluation procedures, including five independent runs, uncertainty quantification, pairwise statistical testing, and equivalence analysis.

- A joint assessment of predictive performance, computational efficiency, and learned feature organization, enabling comparison beyond classification accuracy alone.

- Empirically informed architecture-selection guidance based on the observed trade-offs among accuracy, variability, computational requirements, and feature-space organization.

The remainder of this paper is organized as follows. Section 2 describes the dataset, selected architectures, experimental design, training protocol, and evaluation methodology. Section 3 presents the empirical results across all dimensions of the analysis. Section 4 discusses the principal findings and their implications for architecture selection. Sections 5 addresses study limitations and outlines directions for future research. Section 6 presents the conclusions.

2. MATERIALS AND METHODS

2.1 Overview of the proposed benchmarking framework

This study employs a unified benchmarking framework to enable controlled and reproducible comparisons among the evaluated architectures. All models were trained and evaluated under identical preprocessing, augmentation, optimization, and evaluation protocols. A fixed, stratified, three-way data partition was maintained throughout the experiments, while five independent training runs were conducted with different random seeds. This design allowed for the evaluation of both performance comparisons and variability between runs in a consistent experimental setting.

As shown in Figure 1, the workflow of the proposed framework includes dataset preparation, standardized model training, performance evaluation, statistical and representation-level analysis, and comparative benchmarking. This framework not only measures predictive performance but also evaluates computational efficiency, feature-space representations, statistical significance, performance stability, and accuracy-efficiency trade-offs. This design facilitates cross-paradigm comparisons while assessing the suitability of CNNs and transformer architectures.

2.2 Dataset and preprocessing

The experiments were conducted using the AID and obtained from a publicly available Kaggle repository [22, 23]. The dataset contains 10,000 RGB aerial images distributed across 30 scene categories. The dataset was constructed from Google Earth imagery collected from diverse geographical

regions, seasons, and imaging conditions. Consequently, AID exhibits substantial intra-class variability and inter-class similarity. This makes aerial scene classification challenging. Representative samples from each scene category of the dataset are shown in Figure 2.

The AID dataset was divided using a fixed stratified three-way split, allocating 70% for training, 10% for validation, and 20% for testing. The validation set was used exclusively for model selection, whereas the held-out test set was reserved for

final performance evaluation. The dataset partition was generated using a fixed random seed (42) and reused across all experimental runs to ensure strict comparability among architectures and eliminate test-set leakage. To ensure statistical reliability, five independent training runs were conducted using different random seeds (42-46) while maintaining the same train/validation/test partition throughout the study.

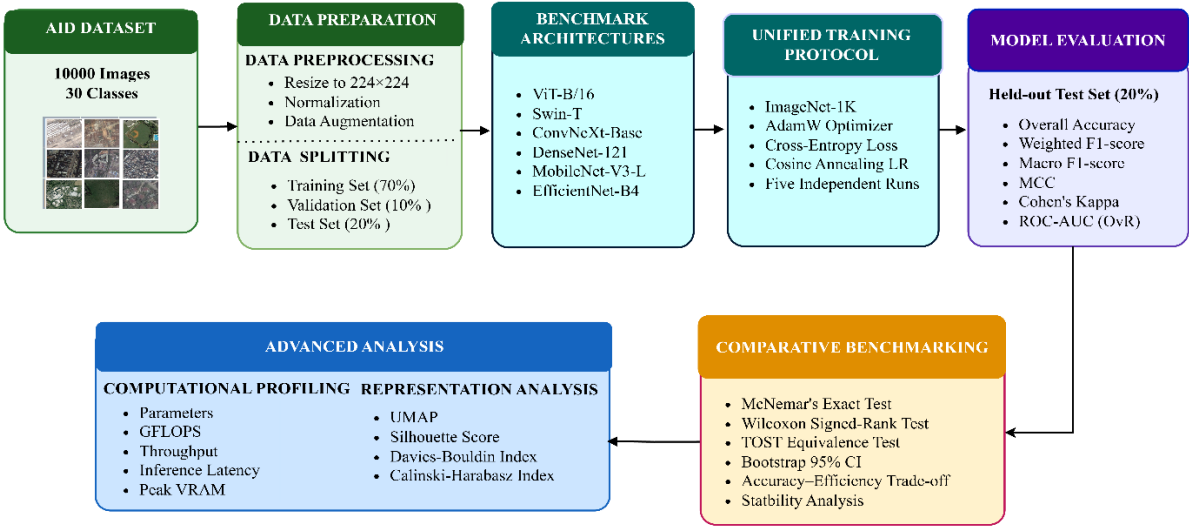


Figure 1. Proposed unified benchmarking framework for aerial scene classification

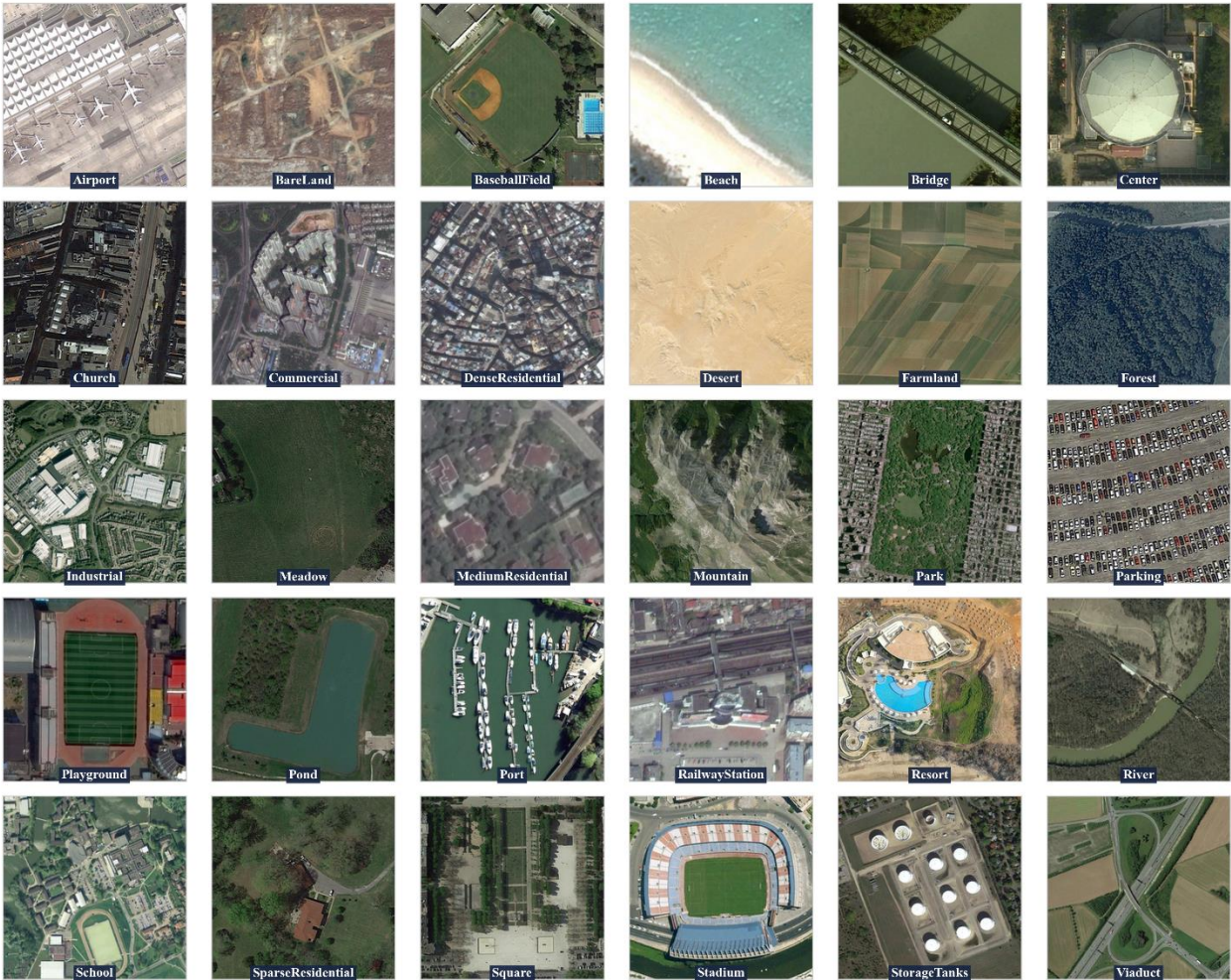


Figure 2. Representative sample images from the 30 scene categories of the Aerial Image Dataset (AID)

Before training, all images were resized from 600×600 pixels to 224×224 pixels to provide a common spatial input size across the evaluated architectures. Inputs were normalized using standard ImageNet-1K channel statistics to match the scaling protocol of the pretrained weights.

The normalization was performed according to Eq. (1):

$$X_{norm} = \frac{X - \mu}{\sigma} \quad (1)$$

where, X denotes the input image tensor, and μ and σ denote the channel-wise mean and standard deviation vectors, respectively. The values used were, $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$. Data augmentation during training included random resized cropping with a scale of 0.8 to 1.0, horizontal and vertical flips ($p = 0.5$), random rotation ($\pm 15^\circ$), color jittering, random grayscale conversion ($p = 0.02$), and random erasing ($p = 0.1$).

2.3 Selection of deep learning architectures

To conduct a rigorous cross-paradigm evaluation, six deep learning architectures representing distinct architectural paradigms were selected.

ViT-B/16 [12] was selected as a representative pure vision transformer. This model partitions images into non-overlapping patches and processes them through stacked self-attention layers. It treats image recognition as a sequence modeling task to incorporate minimal inductive bias towards local spatial structures. (Swin-T) [13] was included as a representative hierarchical transformer architecture. By employing shifted-window self-attention and multi-stage feature extraction, this model reduces computational complexity while maintaining the ability to model long-range contextual relationships.

ConvNeXt-Base [17] was selected to represent modern CNN design. It modernizes the ResNet architecture by

adopting large-kernel depthwise convolutions, LayerNorm, GELU activations, and transformer-inspired design principles. This architecture integrates several design principles inspired by transformer-based models while retaining a fully convolutional structure, thus providing a strong convolutional baseline for comparison. MobileNetV3-L [11] was included to evaluate the performance of lightweight architectures in aerial scene classification. The model combines neural architecture search and hardware-aware optimization to achieve high computational efficiency while maintaining competitive recognition performance [24].

DenseNet-121 [7] was selected as a representative densely connected CNN. Its dense connectivity pattern promotes feature reuse, improves gradient propagation, and enables parameter-efficient learning.

Finally, EfficientNet-B4 was included to evaluate the effectiveness of compound network scaling. Although the original EfficientNet-B4 architecture was designed for a higher input resolution (380×380 pixels), all images were resized to a common resolution of 224×224 pixels in this study.

2.4 Experimental design and standardized training protocol

For each architecture, five independent training runs were conducted using seeds 42-46 while retaining the same fixed train/validation/test partition. For each run, model parameters were updated using the training set, with training samples shuffled at each epoch. The validation set was used for checkpoint selection without parameter updates, whereas the held-out test set was used only for final evaluation using the checkpoint with the highest validation accuracy.

All architectures were trained using the common configuration summarized in Table 1. The same preprocessing, augmentation, optimization, and evaluation protocols were applied across models to ensure comparability.

Table 1. Training configuration of all evaluated models

Hyperparameter	Value/Setting
Dataset & Input	AID (30 classes), images resized to 224×224 pixels
Data Partitioning	Fixed stratified split: 70% training, 10% validation, and 20% held-out testing; partition seed = 42
Experimental Protocol	Five independent runs (seeds 42-46); ImageNet-1K pretrained initialization
	ImageNet Normalization with $\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$
Preprocessing & Augmentation	RandomResizedCrop (scale = 0.8-1.0); horizontal/vertical flips ($p = 0.5$); random rotation ($\pm 15^\circ$); ColorJitter (brightness/contrast/saturation = 0.15, hue = 0.04); RandomGrayscale ($p = 0.02$); RandomErasing ($p = 0.1$); validation/test: resize only, without augmentation
Classification Head	Dropout(0.1) + linear layer with 30 outputs
Loss and Optimization	Class-weighted Cross-Entropy with label smoothing with ϵ of 0.05, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$), Learning rate = 1×10^{-4} , weight decay = 1×10^{-4} ;
Training Strategy	Batch size = 32; 40 maximum epochs; 3-epoch linear warm-up; cosine annealing ($\eta_{min} = 1 \times 10^{-6}$); early stopping (patience = 8, min_delta = 1×10^{-4}), gradient clipping (max-norm = 1.0)
Evaluation & Statistics	ACC, F1-W, F1-M, MCC, Cohen's Kappa, AUC;
Representation analysis	95% confidence intervals, Wilcoxon signed-rank, McNemar, and TOST equivalence tests
Computational profiling	UMAP with Silhouette coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index
	Parameters, GFLOPs, inference latency, peak VRAM, throughput, and training time

Note: Aerial Image Dataset (AID), two one-sided tests (TOST), Uniform Manifold Approximation and Projection (UMAP).

Class-weighted cross-entropy was used as the common loss function across all architectures. Label smoothing distributes a small portion of the target probability mass across non-target classes. The smoothed target distribution was defined as in Eq. (2).

$$q_c = \begin{cases} 1 - \epsilon + \frac{\epsilon}{c}, & c = y \\ \frac{\epsilon}{c}, & c \neq y \end{cases} \quad (2)$$

where, y denotes the ground-truth class label, $C = 30$ represents the number of classes, and $\epsilon = 0.05$ is the label-

smoothing factor.

The optimization objective was defined as in Eq. (3).

$$L = - \sum_{c=1}^C w_c y_c \log(p_c) \quad (3)$$

where, $C = 30$, w_c represents the class weight, y_c represents the ground-truth label, and p_c denotes the predicted probability for class c .

To further stabilize training, gradient clipping was applied during backpropagation. Early stopping was used to limit unnecessary training once validation performance ceased to improve. All experiments were conducted on an NVIDIA GeForce RTX 3060 GPU with 12.9 GB VRAM and implemented in PyTorch. Deterministic CUDA settings were enabled where supported to improve reproducibility.

2.5 Performance evaluation metrics

Model performance was evaluated using multiple complementary classification metrics to provide a comprehensive assessment of predictive capability.

Overall classification accuracy was computed using Eq. (4).

$$Accuracy = \frac{\sum_{i=1}^N I(y_i = \hat{y}_i)}{N} \quad (4)$$

where, N is the total number of samples, y_i is the true label, $\hat{y}_i$ is the predicted label, and $I(\cdot)$ is the indicator function.

Precision and recall were calculated using Eqs. (5) and (6).

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad (5)$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (6)$$

where, TP_c , FP_c and FN_c denote true positives, false positives and false negatives, for class (c), respectively.

The class-wise F1-score was computed using Eq. (7).

$$F1_c = \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \quad (7)$$

Both Macro-F1 and Weighted-F1 scores were reported. Macro-F1 assigns equal importance to all classes, whereas Weighted-F1 accounts for class frequencies. These metrics were calculated according to Eqs. (8) and (9), respectively:

$$F1_{Macro} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (8)$$

$$F1_{Weighted} = \sum_{c=1}^C \frac{n_c}{N} F1_c \quad (9)$$

where, C represents the total number of classes, n_c is the number of samples in class c and N is the total number of samples.

MCC was computed to provide a balanced evaluation of classification performance by considering all elements of the confusion matrix. Cohen's kappa coefficient (κ) was used to quantify agreement between predicted and true labels while accounting for agreement occurring by chance. The coefficient was computed using Eq. (10).

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (10)$$

where, p_0 and p_e denote observed and expected agreement, respectively.

Multiclass ROC-AUC was computed using both One-Versus-Rest (OvR) and One-Versus-One (OvO) strategies. OvR-AUC was used as the primary AUC measure in the comparative results, while OvO-AUC was retained as an additional discrimination measure.

To prevent over-interpreting minor margin differences, pairwise comparisons between architectures were conducted using McNemar's test [25] and the Wilcoxon signed-rank test [26]. McNemar's test was used to evaluate differences in discordant predictions between two architectures on the same held-out test samples.

The Wilcoxon signed-rank test was used for paired comparisons of run-level performance across the five independent runs. Because a non-significant result does not imply practical equivalence, equivalence was additionally assessed using TOST with a pre-specified margin of $\Delta = \pm 0.5$ percentage points. Architectures were considered equivalent only when the corresponding 90% confidence interval for the paired mean accuracy difference was entirely contained within $[-\Delta, +\Delta]$. Accordingly, non-significant McNemar or Wilcoxon results are reported as "no significant difference detected" and are not interpreted as evidence of equivalence unless the TOST criterion is also satisfied.

To quantify uncertainty in the reported test-set metrics, 2,000 bootstrap resamples of the 2,007 test predictions were generated. The 95% confidence interval was estimated using the percentile bootstrap method according to Eq. (11).

$$CI_{95\%} = [\theta_{0.025}^*, \theta_{0.975}^*] \quad (11)$$

where, $\theta_{0.025}^*$ and $\theta_{0.975}^*$ represent the lower and upper percentiles of the bootstrap distribution, respectively.

Feature-space characteristics were analysed using UMAP [27]. Penultimate-layer feature vectors were extracted from all 2,007 test images and L2-normalized before projection into a two-dimensional embedding space. UMAP was configured with `n_neighbors = 30`, `min_dist = 0.10`, the Euclidean distance metric, and `random_state = 42`. Cluster quality was quantitatively assessed using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. All test samples were visualized with an alpha value of 0.45 to reduce overplotting. The quantitative cluster-separation metrics were reported alongside the UMAP visualizations to facilitate a comprehensive assessment of feature-space separability across the evaluated architectures.

Operational deployment readiness was tracked using five computational indicators, such as trainable parameter count, computational complexity (GFLOPs), inference latency, inference throughput, and peak VRAM memory consumption.

3. RESULTS

3.1 Optimization trajectories and convergence dynamics

Figure 3 presents the training and validation learning curves for all six architectures from Run 0 with random seed 42.

ViT-B/16 exhibited rapid initial learning, with validation accuracy increasing substantially during the early epochs

before reaching a plateau. The best checkpoint was obtained at epoch 14 with validation accuracy of 0.9590. The learning curves also showed a relatively larger training-validation gap

than those of several other architectures. Its corresponding Run-0 test accuracy was 95.57%.

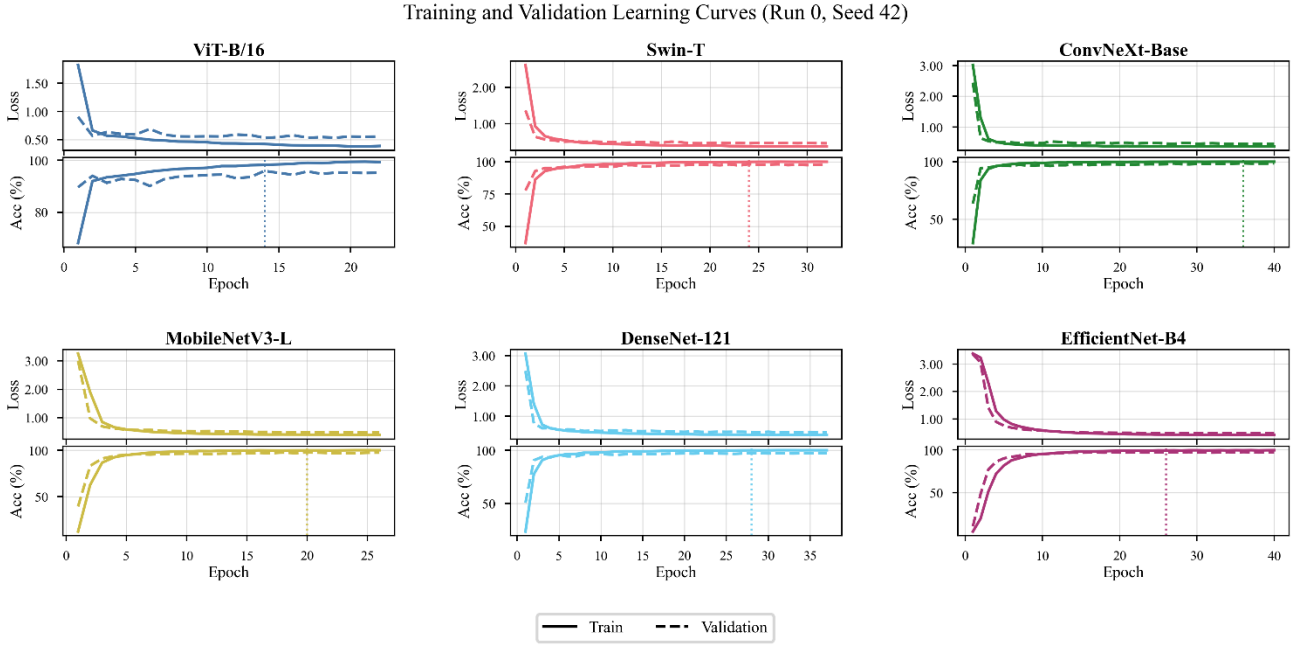


Figure 3. Training and validation loss and accuracy curves for the six evaluated architectures in Run 0 (seed 42) on Aerial Image Dataset (AID)

Swin-T showed a relatively smooth validation trajectory, with progressive improvement and limited epoch-to-epoch fluctuation. The best checkpoint was obtained at epoch 24, with a validation accuracy of 0.9800. Its Run-0 test accuracy was 97.96%, while its five-run mean test accuracy was 97.88%.

ConvNeXt-Base showed slower initial adaptation but continued to improve for a longer portion of the training process. The model achieved its best checkpoint at epoch 36 with a validation accuracy of 0.9850, marking the latest checkpoint among all architectures. Its corresponding Run-0 test accuracy was 98.31%.

MobileNetV3-Large showed rapid early improvement followed by stabilization of validation performance. Its best checkpoint was obtained at epoch 20, with a validation accuracy of 0.9760. Despite its substantially smaller parameter count, the model achieved a Run-0 test accuracy of 97.46%.

DenseNet-121 and EfficientNet-B4 exhibited similar optimization trajectories characterized by rapid early improvements followed by gradual refinement. Their best checkpoints were obtained at epochs 28 with validation accuracy of 0.9730 and 26 validation accuracy of 0.9740, respectively. Both architectures achieved stable convergence and competitive final accuracies, although neither matched the performance of Swin-T or ConvNeXt-Base.

Overall, the learning curves indicate that all architectures converged within 40 epochs. Swin-T achieved the strongest combination of convergence stability and predictive performance, whereas MobileNetV3-L offered the most favorable balance between accuracy and computational efficiency.

3.2 Classification performance

Table 2 reports the complete run-level classification results

obtained on the held-out AID test set across five independent runs (seeds 42-46), while Table 3 provides the corresponding five-run mean $\pm$ standard deviation for direct comparison among the six architectures. We report both levels of results to make the evaluation transparent while distinguishing aggregate performance from run-to-run variability.

Across the five-run evaluation, mean test accuracy ranged from 96.38% to 97.88% (Table 3). Swin-T achieved the highest mean accuracy ($97.88 \pm 0.16\%$), followed by ConvNeXt-Base ($97.65 \pm 0.48\%$) and MobileNetV3-L ($97.51 \pm 0.19\%$). The corresponding weighted F1-scores closely followed the accuracy results, with Swin-T obtaining the highest value ($97.88 \pm 0.15\%$). Macro-F1 showed the same overall ordering, ranging from $96.25 \pm 1.34\%$ for ViT-B/16 to $97.76 \pm 0.16\%$ for Swin-T.

The run-level results in Table 2 provide additional evidence of differences in training stability. ViT-B/16 exhibited the widest accuracy range, from 94.47% to 97.36%, corresponding to the largest standard deviation (1.31 percentage points). In contrast, Swin-T varied only from 97.71% to 98.11%, with an SD of 0.16 percentage points. MobileNetV3-L also showed low variability (0.19 percentage points), whereas ConvNeXt-Base exhibited greater variation (0.48 percentage points). These results indicate greater sensitivity of ViT-B/16 performance to the evaluated random seeds under the adopted training protocol.

The agreement among the classification metrics further supports the observed performance pattern. All six architectures achieved very high one-vs-rest ROC-AUC, ranging from 0.9985 ± 0.0005 to 0.9996 ± 0.0001 . EfficientNet-B4 recorded the highest mean AUC-OvR (0.9996 ± 0.0001), despite its lower mean accuracy than Swin-T. Thus, the high AUC values indicate strong multiclass discrimination across the evaluated architectures, but they do not by themselves establish superiority in overall classification

accuracy.

Figure 4 presents the five-run mean test accuracy with standard-deviation error bars, providing a compact visualization of the consolidated performance reported in Table 3. The figure highlights the narrow variability of Swin-

T and MobileNetV3-L compared with the greater run-to-run variability observed for ViT-B/16. The corresponding training and validation loss and accuracy trajectories are presented separately in Figure 3.

Table 2. Run-level classification performance of the evaluated architectures on the held-out Aerial Image Dataset (AID) test set across five independent runs

Model	Run	Seed	Accuracy (%)	F1-W (%)	F1-M (%)	MCC	Kappa	AUC-OvR	Best Epoch	Best Val. Accuracy (%)	Best Val. Loss
ViT-B/16	0	42	95.57	95.56	95.43	0.954146	0.954080	0.9982	14	95.9	0.5334
ViT-B/16	1	43	97.31	97.31	97.19	0.972153	0.972138	0.9982	38	97.4	0.4779
ViT-B/16	2	44	94.47	94.47	94.27	0.942814	0.942731	0.9990	5	94.9	0.5320
ViT-B/16	3	45	97.21	97.17	97.08	0.971134	0.971107	0.9990	35	97.0	0.4978
ViT-B/16	4	46	97.36	97.35	97.28	0.972665	0.972654	0.9980	35	96.7	0.5192
Swin-T	0	42	97.96	97.96	97.87	0.978862	0.978846	0.9995	24	98.0	0.4546
Swin-T	1	43	97.81	97.80	97.66	0.977316	0.977297	0.9993	31	98.4	0.4430
Swin-T	2	44	97.81	97.80	97.68	0.977315	0.977297	0.9987	22	97.8	0.4608
Swin-T	3	45	98.11	98.10	97.98	0.980398	0.980393	0.9997	34	98.4	0.4409
Swin-T	4	46	97.71	97.73	97.62	0.976295	0.976267	0.9995	15	97.9	0.4521
ConvNeXt-Base	0	42	98.31	98.30	98.23	0.982464	0.982457	0.9991	36	98.5	0.4500
ConvNeXt-Base	1	43	97.16	97.16	96.97	0.970635	0.970590	0.9982	22	97.6	0.4873
ConvNeXt-Base	2	44	97.96	97.95	97.91	0.978855	0.978845	0.9993	29	98.6	0.4450
ConvNeXt-Base	3	45	97.26	97.23	97.15	0.971653	0.971621	0.9982	17	97.8	0.4751
ConvNeXt-Base	4	46	97.56	97.55	97.49	0.974750	0.974717	0.9991	17	97.8	0.4698
MobileNetV3-L	0	42	97.46	97.46	97.35	0.973704	0.973685	0.9991	20	97.6	0.4796
MobileNetV3-L	1	43	97.21	97.21	97.12	0.971121	0.971104	0.9990	27	97.5	0.4802
MobileNetV3-L	2	44	97.61	97.60	97.44	0.975243	0.975233	0.9995	35	97.7	0.4579
MobileNetV3-L	3	45	97.71	97.70	97.63	0.976274	0.976265	0.9990	21	96.9	0.4888
MobileNetV3-L	4	46	97.56	97.55	97.46	0.974731	0.974717	0.9992	30	97.6	0.4746
DenseNet-121	0	42	96.76	96.71	96.58	0.966498	0.966460	0.9996	28	97.3	0.4725
DenseNet-121	1	43	97.16	97.16	97.03	0.970618	0.970590	0.9991	29	97.8	0.4582
DenseNet-121	2	44	97.71	97.69	97.60	0.976282	0.976265	0.9996	33	97.6	0.4588
DenseNet-121	3	45	97.26	97.26	97.10	0.971638	0.971621	0.9996	27	98.0	0.4521
DenseNet-121	4	46	96.91	96.89	96.79	0.968036	0.968010	0.9992	37	97.8	0.4486
EfficientNet-B4	0	42	96.56	96.55	96.41	0.964426	0.964397	0.9995	26	97.4	0.4778
EfficientNet-B4	1	43	96.71	96.70	96.56	0.965972	0.965945	0.9996	23	96.4	0.4891
EfficientNet-B4	2	44	97.06	97.04	96.93	0.969574	0.969558	0.9996	39	97.0	0.4801
EfficientNet-B4	3	45	97.06	97.04	96.92	0.969576	0.969557	0.9997	31	97.2	0.4815
EfficientNet-B4	4	46	97.21	97.18	97.07	0.971123	0.971106	0.9997	37	96.7	0.4759

Table 3. Five-run mean $\pm$ SD classification performance of the evaluated architectures on the held-out Aerial Image Dataset (AID) test set

Model	ACC (%)	F1-W (%)	F1-M (%)	MCC	Kappa	AUC-OvR	Params (M)
ViT-B/16	96.38 $\pm$ 1.31	96.37 $\pm$ 1.30	96.25 $\pm$ 1.34	0.9626 $\pm$ 0.0135	0.9625 $\pm$ 0.0135	0.9985 $\pm$ 0.0005	85.82
Swin-T	97.88 $\pm$ 0.16	97.88 $\pm$ 0.15	97.76 $\pm$ 0.16	0.9780 $\pm$ 0.0016	0.9780 $\pm$ 0.0016	0.9993 $\pm$ 0.0004	27.54
ConvNeXt-Base	97.65 $\pm$ 0.48	97.64 $\pm$ 0.49	97.55 $\pm$ 0.52	0.9757 $\pm$ 0.0050	0.9756 $\pm$ 0.0050	0.9988 $\pm$ 0.0005	87.60
MobileNetV3-L	97.51 $\pm$ 0.19	97.50 $\pm$ 0.19	97.40 $\pm$ 0.19	0.9742 $\pm$ 0.0020	0.9742 $\pm$ 0.0020	0.9991 $\pm$ 0.0002	4.24
DenseNet-121	97.16 $\pm$ 0.36	97.14 $\pm$ 0.37	97.02 $\pm$ 0.38	0.9706 $\pm$ 0.0038	0.9706 $\pm$ 0.0038	0.9994 $\pm$ 0.0002	6.98
EfficientNet-B4	96.92 $\pm$ 0.27	96.90 $\pm$ 0.26	96.78 $\pm$ 0.28	0.9681 $\pm$ 0.0028	0.9681 $\pm$ 0.0028	0.9996 $\pm$ 0.0001	17.60

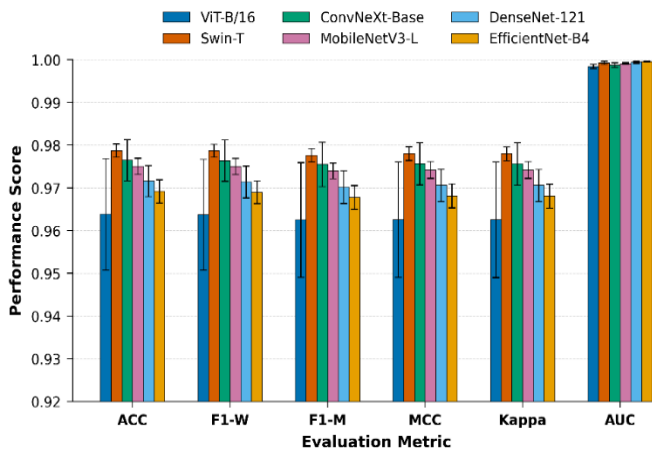


Figure 4. Five-run mean test accuracy of the evaluated architectures with standard-deviation error bars

Overall, Swin-T provided the highest mean classification performance under the adopted experimental protocol, while ViT-B/16 showed the greatest run-to-run variability. However, the relatively small differences among the stronger-performing architectures should not be interpreted as definitive evidence of superiority based solely on mean accuracy. Their statistical relationships are examined in the subsequent pairwise significance and equivalence analyses.

3.3 Mean per-class F1 analysis

Per-class analysis in run 0 revealed that the high aggregate classification performance was accompanied by substantial variation across individual scene categories (Figure 5). Several visually distinctive classes, including Beach, Mountain, Forest, BaseballField, and Parking, achieved F1-scores close to 1.00 across most architectures. In contrast, lower F1-scores were observed for a smaller group of more challenging

categories, particularly Church, Square, Park, School, and Stadium. The differences among architectures were therefore more apparent at the class level than in the aggregate accuracy measures. Swin-T showed relatively consistent performance across several of the lower-performing classes, although class-specific strengths and weaknesses were observed for all architectures. These results indicate that the small differences in overall performance arise partly from the handling of a limited set of visually challenging scene categories rather than from uniformly different performance across all classes.

3.4 Confusion matrix analysis

Figure 6 presents the mean row-normalized confusion matrices across five independent runs on the held-out AID test set. The matrices show a predominantly strong diagonal pattern across all six architectures, consistent with their high aggregate test performance.

Class-wise recall, however, varies across categories. Resort, School, Park, and Square exhibit comparatively lower diagonal values, whereas several classes attain recall close to 1.00 across most architectures. The off-diagonal values remain generally small, with their distribution varying across models and classes.

Swin-T, which achieved the highest mean test accuracy (97.88%), also shows high diagonal values across most classes. Nevertheless, the matrices indicate that high aggregate accuracy does not correspond to uniform performance across all scene categories.

Overall, the five-run confusion-matrix analysis

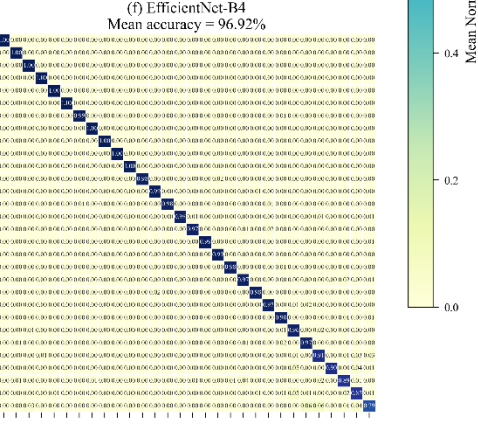
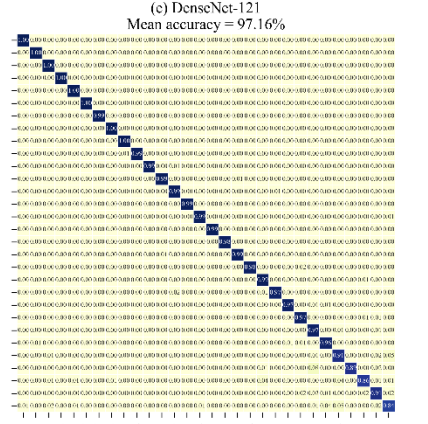
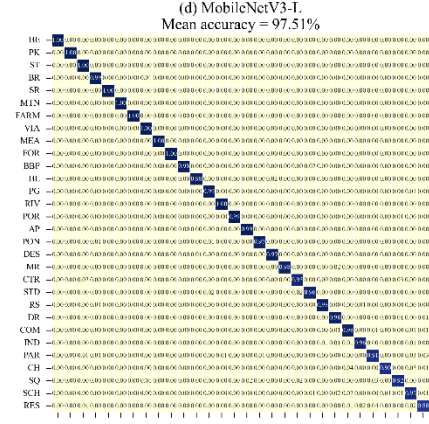
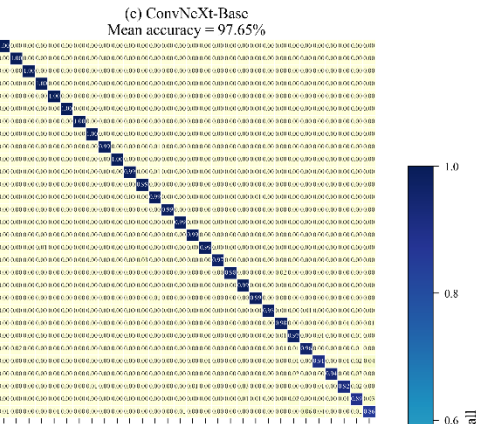
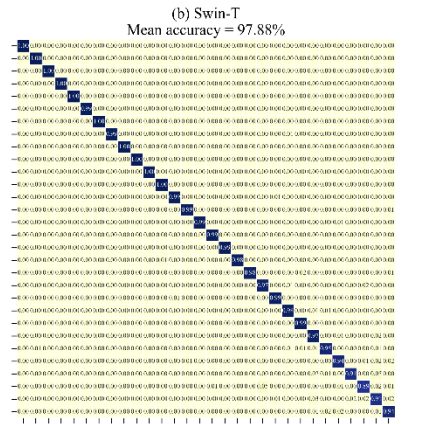
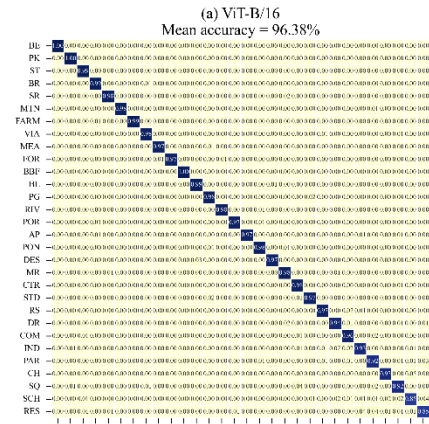


Figure 6. Mean row-normalized confusion matrices across five independent runs (seeds 42-46) on the held-out Aerial Image Dataset (AID) test set

complements the aggregate and per-class metrics by revealing where classification performance varies across the 30 AID classes, while avoiding conclusions beyond the observed prediction patterns.

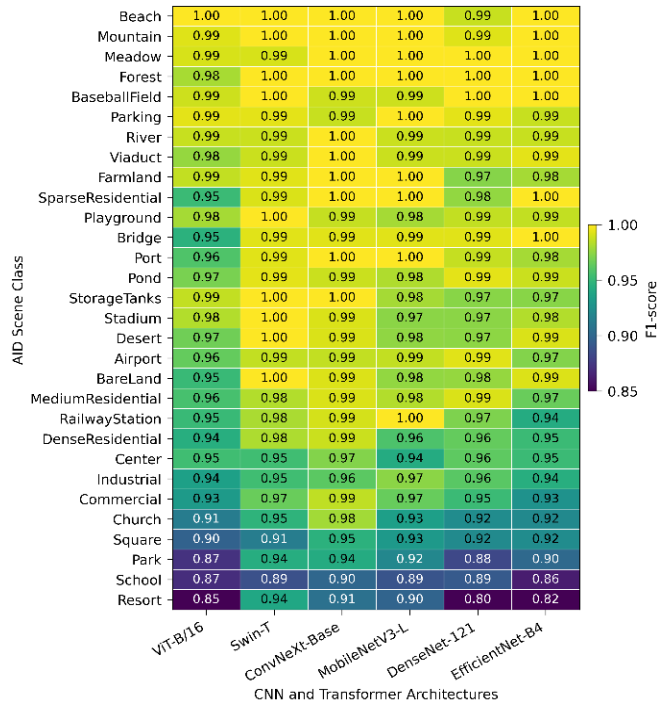


Figure 5. Per-class F1-score heatmap

3.5 Accuracy-complexity trade-off

The computational efficiency comparisons of the evaluated architectures under identical experimental conditions are summarized in Table 4. Swin-T achieved the highest mean test accuracy ($97.88 \pm 0.16\%$) while requiring substantially fewer parameters than ViT-B/16 and ConvNeXt-Base. In contrast, ConvNeXt-Base and ViT-B/16 had parameter counts exceeding 85 M without achieving higher mean accuracy than Swin-T. These results indicate that increasing model size alone did not provide a corresponding improvement in predictive performance under the adopted experimental protocol.

MobileNetV3-L exhibited the most compact computational profile, with only 4.24 M parameters and 0.23 GFLOPs, while maintaining a competitive mean accuracy of $97.51 \pm 0.19\%$. It also recorded the lowest measured inference latency (4.66 ± 0.51 ms/image). Thus, MobileNetV3-L provides a particularly favorable accuracy-efficiency profile when computational and latency constraints are considered.

Inference latency did not follow a simple relationship with either parameter count or GFLOPs. For example, Swin-T required higher inference latency than ViT-B/16 despite having substantially fewer parameters, illustrating that parameter count and theoretical computational complexity do not fully determine observed execution speed. Measured latency is additionally influenced by the hardware, implementation, memory access, and execution configuration.

As illustrated in Figure 7, the evaluated architectures therefore occupy different positions in the accuracy–complexity space. Swin-T provides the highest predictive performance with a substantially smaller parameter footprint than the larger ViT-B/16 and ConvNeXt-Base models, whereas MobileNetV3-L offers the smallest computational footprint and lowest measured latency while retaining competitive accuracy. These findings highlight the importance of jointly considering predictive performance and computational characteristics when selecting architectures for aerial scene classification.

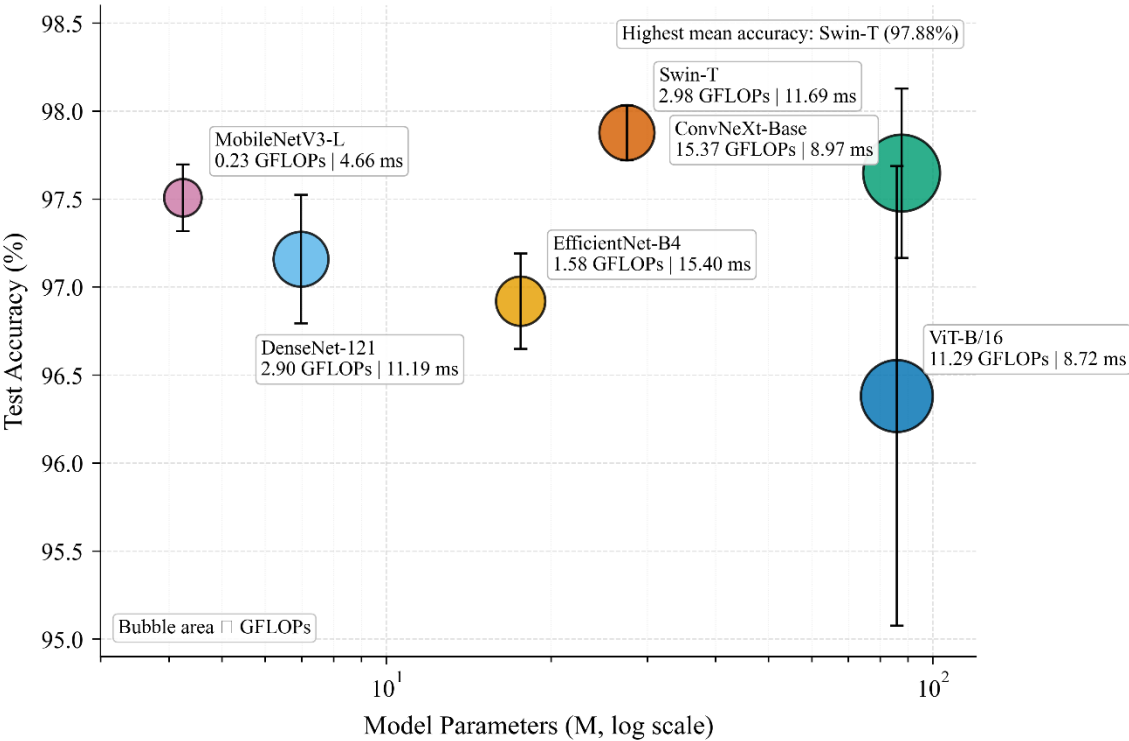


Figure 7. Accuracy-complexity trade-off among the evaluated architectures

Table 4. Computational efficiency comparisons of the evaluated architectures under identical experimental conditions

Model	Parameters (M)	GFLOPs	Latency ms/Image	Peak VRAM MB	Accuracy (%)
ViT-B/16	85.82	11.29	8.72 ± 0.06	2877 ± 180	96.38 ± 1.31
Swin-T	27.54	2.98	11.69 ± 3.05	2918 ± 143	97.88 ± 0.16
ConvNeXt-Base	87.60	15.37	8.97 ± 0.03	3537 ± 184	97.65 ± 0.48
MobileNetV3-L	4.24	0.23	4.66 ± 0.51	3101 ± 0	97.51 ± 0.19
DenseNet-121	6.98	2.90	11.19 ± 0.16	3230 ± 12	97.16 ± 0.36
EfficientNet-B4	17.60	1.58	15.40 ± 3.15	2719 ± 459	96.92 ± 0.27

3.6 Statistical significance and equivalence analysis

Pairwise statistical analyses were conducted across the six evaluated architectures using the five independent training runs. McNemar's exact test was applied separately to the held-out test predictions from each run, and the resulting five run-specific p-values were summarized using their median. The Wilcoxon signed-rank test was applied to the five paired run-

level accuracy values. Statistical significance was assessed at $\alpha = 0.05$, and the resulting p-values are presented in Figure 8.

The median McNemar p-values were 0.020 for ViT-B/16 versus Swin-T and 0.009 for Swin-T versus EfficientNet-B4. The remaining pairwise comparisons had median McNemar p-values above 0.05. The Wilcoxon signed-rank test produced p-values ranging from 0.0625 to 0.875 across the 15 pairwise comparisons, with none below 0.05 threshold.

A TOST procedure was also performed on the five paired run-level accuracy differences using a predefined equivalence margin of ± 0.5 percentage points (Figure 9). None of the 15 pairwise comparisons satisfied the two one-sided significance requirements for equivalence. Thus, equivalence within the predefined accuracy margin could not be established for any model pair.

Overall, the analysis showed limited evidence of consistent

pairwise differences. The Wilcoxon analysis did not identify statistically significant differences in run-level accuracy at $\alpha = 0.05$, while the median of the run-specific McNemar tests was below 0.05 for two model pairs. The absence of statistical significance was not interpreted as evidence of equivalence, as none of the comparisons satisfied the predefined TOST criterion.

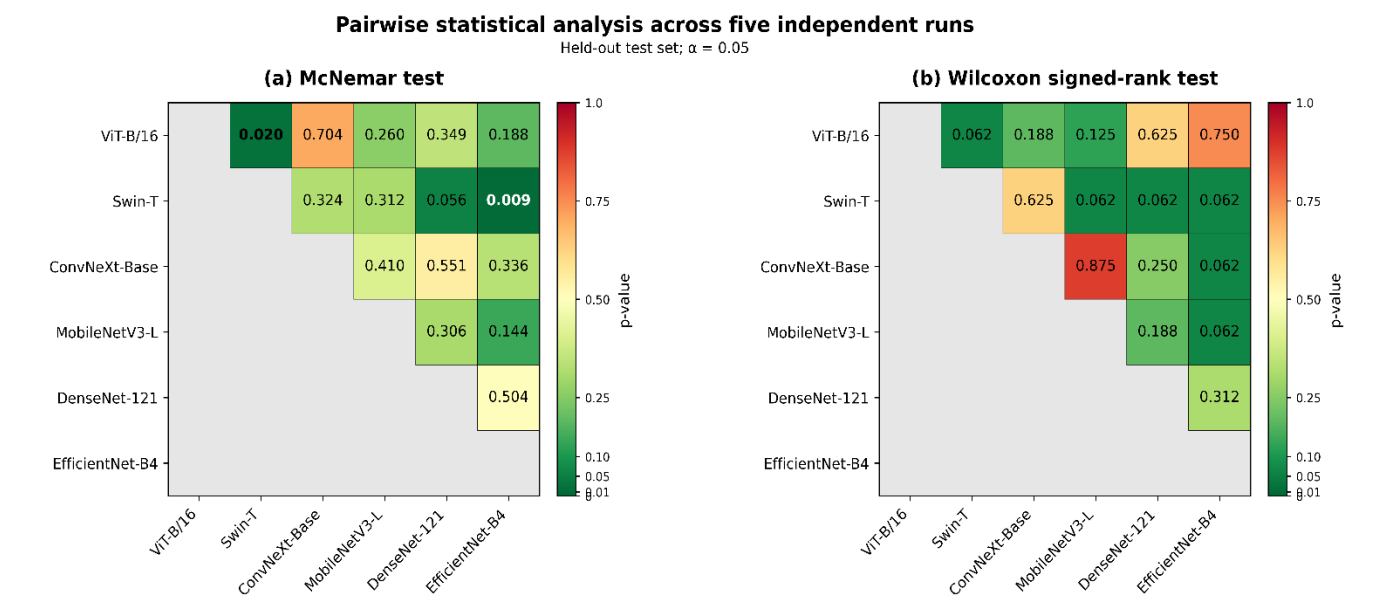


Figure 8. Pairwise statistical analysis of the evaluated architectures. (a) Median p-values from run-specific McNemar exact tests on the held-out test set. (b) Wilcoxon signed-rank p-values based on five paired run-level accuracy values

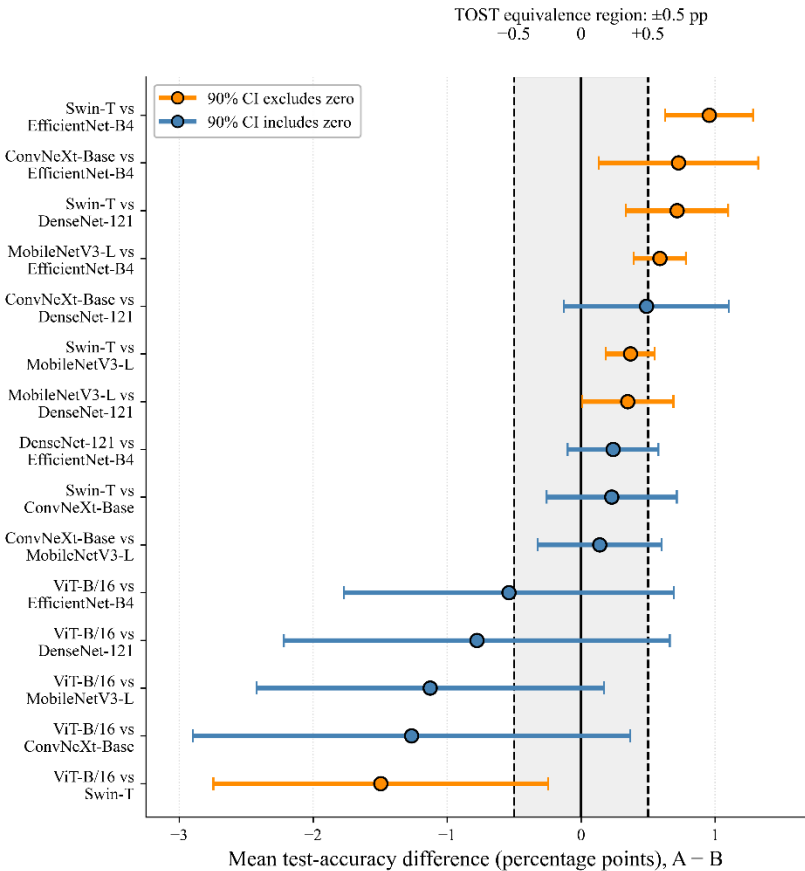


Figure 9. Pairwise test-accuracy differences with 90% CIs and the two one-sided tests (TOST) equivalence margin of ± 0.5 percentage points

3.7 Bootstrap confidence intervals

To assess the stability of the reported performance across repeated training, 2,000 bootstrap resamples were generated from the five independent run-level estimates for each architecture. Percentile-based 95% bootstrap intervals were obtained from the 2.5th and 97.5th percentiles of the resulting bootstrap distributions. Accuracy, macro-averaged F1-score, Matthews correlation coefficient (MCC), Cohen’s κ , and one-vs-rest ROC-AUC were considered.

The results in Table 5 show relatively consistent performance across the five runs for most architectures. Swin-T achieved the highest mean accuracy of 0.9788 ± 0.0016 , with a 95% bootstrap interval of $[0.9777, 0.9800]$. ConvNeXt-Base achieved 0.9765 ± 0.0048 $[0.9728, 0.9802]$, while MobileNetV3-L achieved 0.9751 ± 0.0019 $[0.9735, 0.9764]$. DenseNet-121 and EfficientNet-B4 obtained mean accuracies of 0.9716 ± 0.0036 $[0.9690, 0.9746]$ and 0.9692 ± 0.0027 $[0.9669, 0.9712]$, respectively. ViT-B/16 exhibited comparatively greater run-to-run variation, with an accuracy of 0.9638 ± 0.0131 and a corresponding interval of $[0.9527, 0.9731]$.

The macro-F1 and MCC results followed the same general ordering, with Swin-T providing the highest mean values among the evaluated architectures. ROC-AUC (OvR) remained high for all six models, ranging from 0.9985 to 0.9996, with small between-run standard deviations. Thus, the repeated-run analysis supports the consistency of the observed performance patterns.

Because the bootstrap distributions are based on only five independent training runs, the resulting intervals are interpreted as measures of run-to-run experimental uncertainty rather than as population-level confidence bounds. Formal claims of statistical superiority are therefore based on the dedicated pairwise statistical analyses rather than on confidence-interval overlap alone.

Overall, ViT-B/16 exhibited the greatest uncertainty across the evaluated metrics, whereas the remaining architectures showed substantially smaller run-to-run variability. The

bootstrap analysis therefore complements the repeated-run results by quantifying uncertainty around the observed performance estimates without being used as a substitute for formal pairwise significance or equivalence testing.

3.8 Uniform Manifold Approximation and Projection feature embedding analysis

To examine differences in the learned feature representations, UMAP was applied to penultimate-layer features extracted from the same 2,007 held-out AID test samples for all six architectures [27]. The extracted features were L2-normalized before dimensionality reduction. Figure 10 presents the resulting two-dimensional embeddings, with colors denoting the 30 scene classes.

The embeddings exhibited varying degrees of class compactness and separation across the evaluated architectures (Figure 10). Quantitative assessment of the extracted feature representations further revealed differences in class separability (Table 6). ConvNeXt-Base achieved the highest Silhouette coefficient (0.8845), followed by Swin-T (0.8664) and MobileNetV3-L (0.8526), whereas ViT-B/16 obtained the lowest value (0.7787). For the Davies–Bouldin Index, where lower values indicate more compact and separated clusters, MobileNetV3-L achieved the lowest value (0.2609), followed by EfficientNet-B4 (0.3363) and ConvNeXt-Base (0.3488). The Calinski–Harabasz Index showed another ordering, with EfficientNet-B4 recording the highest value (5,847.66), followed by MobileNetV3-L (5,274.93) and DenseNet-121 (3,658.80) (Table 6).

The three indices therefore do not yield a uniform ranking across architectures (Table 6). ConvNeXt-Base ranks highest by the Silhouette coefficient, MobileNetV3-L by the Davies–Bouldin Index, and EfficientNet-B4 by the Calinski–Harabasz Index. This variation indicates that the architectures differ in multiple aspects of feature-space organization and that no single separability measure provides a complete characterization of the learned representations.

Table 5. Bootstrap mean estimates and 95% confidence intervals obtained from 2,000 bootstrap resamples

Model	Accuracy	F1-macro	MCC	Cohen’s κ	ROC-AUC (OvR)
ViT-B/16	0.9638 ± 0.0131 [0.9527, 0.9731]	0.9625 ± 0.0134 [0.9511, 0.9721]	0.9626 ± 0.0135 [0.9511, 0.9722]	0.9625 ± 0.0135 [0.9510, 0.9721]	0.9985 ± 0.0005 [0.9981, 0.9988]
Swin-T	0.9788 ± 0.0016 [0.9777, 0.9800]	0.9776 ± 0.0016 [0.9765, 0.9789]	0.9780 ± 0.0016 [0.9769, 0.9793]	0.9780 ± 0.0016 [0.9769, 0.9793]	0.9993 ± 0.0004 [0.9990, 0.9996]
ConvNeXt-Base	0.9765 ± 0.0048 [0.9728, 0.9802]	0.9755 ± 0.0052 [0.9715, 0.9795]	0.9757 ± 0.0050 [0.9719, 0.9795]	0.9756 ± 0.0050 [0.9718, 0.9795]	0.9988 ± 0.0005 [0.9984, 0.9992]
MobileNetV3-L	0.9751 ± 0.0019 [0.9735, 0.9764]	0.9740 ± 0.0019 [0.9725, 0.9754]	0.9742 ± 0.0020 [0.9726, 0.9756]	0.9742 ± 0.0020 [0.9725, 0.9755]	0.9991 ± 0.0002 [0.9990, 0.9993]
DenseNet-121	0.9716 ± 0.0036 [0.9690, 0.9746]	0.9702 ± 0.0038 [0.9675, 0.9734]	0.9706 ± 0.0038 [0.9679, 0.9737]	0.9706 ± 0.0038 [0.9679, 0.9737]	0.9994 ± 0.0002 [0.9992, 0.9996]
EfficientNet-B4	0.9692 ± 0.0027 [0.9669, 0.9712]	0.9678 ± 0.0028 [0.9654, 0.9698]	0.9681 ± 0.0028 [0.9658, 0.9702]	0.9681 ± 0.0028 [0.9657, 0.9702]	0.9996 ± 0.0001 [0.9996, 0.9997]

Table 6. Quantitative assessment of learned feature separability on the Aerial Image Dataset (AID) test set

Model	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harabasz Index
ViT-B/16	0.7787	0.6180	1,746.16
Swin-T	0.8664	0.4001	3,291.35
ConvNeXt-Base	0.8845	0.3488	2,781.73
MobileNetV3-L	0.8526	0.2609	5,274.93
DenseNet-121	0.8004	0.4602	3,658.80
EfficientNet-B4	0.8235	0.3363	5,847.66

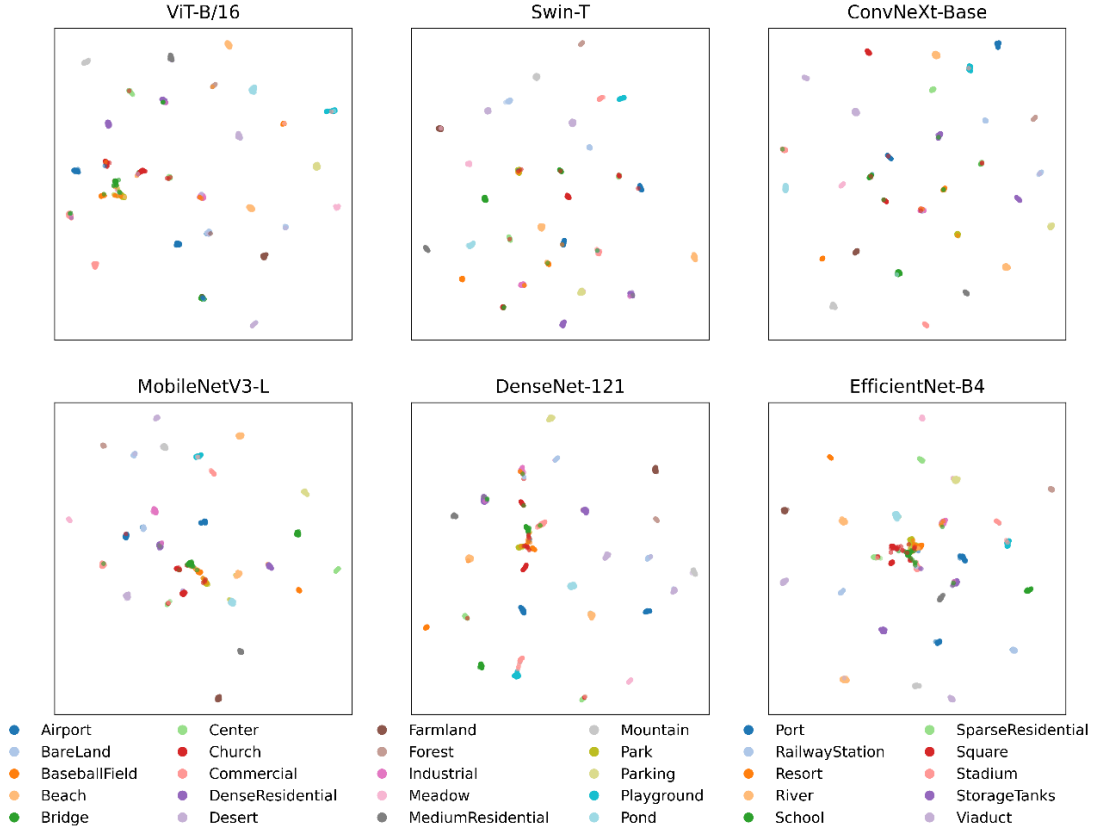


Figure 10. Uniform Manifold Approximation and Projection (UMAP) visualization of learned aerial-scene representations on the Aerial Image Dataset (AID) test set

The representation analysis complements the classification and computational results. In particular, MobileNetV3-L combines relatively favorable feature-space measurements with a substantially smaller computational footprint, whereas ConvNeXt-Base achieves the highest Silhouette coefficient at considerably higher parameter and FLOP requirements. These observations provide an additional perspective on the architectural differences identified through classification performance and computational profiling.

4. DISCUSSION

The five-run evaluation shows that the six architectures achieve consistently high performance on AID, but with clear differences in model size and computational cost. Swin-T obtained the highest mean accuracy ($97.88 \pm 0.16\%$) and weighted F1 ($97.88 \pm 0.15\%$), followed closely by ConvNeXt-Base ($97.65 \pm 0.48\%$) and MobileNetV3-L ($97.51 \pm 0.19\%$). The relatively small standard deviations for these models indicate stable performance across repeated runs.

The results also indicate that higher classification performance is not directly associated with greater model complexity. ConvNeXt-Base contains 87.60 M parameters and requires 15.37 GFLOPs, whereas Swin-T uses 27.54 M parameters and 2.98 GFLOPs while achieving higher accuracy. MobileNetV3-L provides another efficient operating point, with only 4.24 M parameters and 0.23 GFLOPs, while maintaining $97.51 \pm 0.19\%$ accuracy. Thus, the comparison identifies meaningful differences in the accuracy-complexity relationship rather than a simple performance advantage for larger architectures.

The per-class analysis further indicates that the high

aggregate scores do not imply uniform discrimination across all AID scene categories. Differences in class-wise F1 and the normalized confusion matrices indicate that some scene categories remain more difficult to distinguish than others. This is particularly relevant for aerial scene classification because visually related categories can share structural and spatial characteristics. The class-level results therefore complement the aggregate metrics by identifying where classification errors are concentrated.

The statistical analyses provide additional context for interpreting the small differences between the stronger models. Although several models achieved similar mean performance, the pairwise significance and equivalence analyses indicate that statistical similarity and practical equivalence should not be inferred from mean accuracy or confidence-interval overlap alone. Consequently, model selection should consider predictive performance jointly with computational complexity, latency, memory requirements, and deployment constraints.

Overall, the experiments suggest that Swin-T provides the strongest balance between classification performance and computational cost among the evaluated architectures, while MobileNetV3-L represents a substantially lighter alternative with only a modest reduction in accuracy. These conclusions are specific to the adopted training protocol, AID test set, and computational environment and should not be interpreted as universal architectural rankings.

5. LIMITATIONS AND FUTURE WORK

Several limitations should be considered when interpreting the findings of this study.

The study has four main limitations. First, the evaluation is restricted to the AID dataset and its 30 scene categories. The observed differences among architectures may therefore be influenced by the characteristics of this dataset and may not be directly transferable to other aerial datasets or acquisition conditions.

Second, the use of five independent runs captures variability associated with repeated training, but does not assess performance under alternative data partitions or distribution shifts. The behaviour of the evaluated models under changes in geographic region, sensor characteristics, spatial resolution, illumination, or scene composition was not examined.

Third, the computational measurements are dependent on the experimental hardware and implementation. While parameter count and GFLOPs provide architecture-level measures, latency, throughput, and GPU memory depend on the hardware, input size, batch configuration, and software environment. The reported measurements should therefore be interpreted as comparative results under the conditions used in this study.

Fourth, the analysis identifies differences in class-level performance through F1 scores and confusion matrices, but does not directly investigate the visual or semantic characteristics associated with individual classification errors. In addition, the comparison considers six selected architectures under a common training protocol and does not cover the broader range of models available for aerial scene classification.

These limitations define the scope of the reported findings. Further evaluation using additional datasets, alternative data distributions, and different computational platforms would help determine whether the observed performance and efficiency patterns remain consistent beyond the experimental setting considered here.

Future research should extend the proposed benchmarking framework beyond the AID dataset to evaluate the robustness and generalizability of the observed architecture rankings across diverse remote sensing benchmarks. This work can be extended with nested k-fold cross-validation across multiple partition boundaries to directly test whether the reported accuracies reflect generalization to unseen regions. In particular, the influence of domain-specific and self-supervised pretraining strategies require further investigation, as recent advances in remote sensing foundation models may alter the relative performance of transformer-based architectures. Additional studies should also explore architecture-specific optimization strategies, model compression techniques, and knowledge distillation framework to further improve deployment efficiency. Finally, extending the evaluation to uncertainty-aware learning, interpretability analysis, UAV-acquired and temporal imagery, and other remote sensing tasks would provide a more comprehensive understanding of architectural suitability for operational geospatial applications.

6. CONCLUSION

This study presents a controlled comparison of six CNN and Transformer architectures for 30-class aerial scene classification on the AID dataset. Under a common training and evaluation protocol with five independent runs, all models achieved high classification performance, with mean test accuracy ranging from $96.38 \pm 1.31\%$ to $97.88 \pm 0.16\%$. Swin-

T achieved the highest mean accuracy ($97.88 \pm 0.16\%$) and weighted F1 ($97.88 \pm 0.15\%$) among the evaluated models.

The comparison also shows substantial differences in computational requirements. Swin-T used 27.54 M parameters and 2.98 GFLOPs, compared with 87.60 M parameters and 15.37 GFLOPs for ConvNeXt-Base, despite their similar classification performance. MobileNetV3-L achieved $97.51 \pm 0.19\%$ accuracy with only 4.24 M parameters and 0.23 GFLOPs, representing a substantially smaller computational footprint.

The repeated-run results indicate that performance differences should be considered together with run-to-run variability and the statistical analyses. The consistently high AUC-OvR values (0.9985-0.9996) were accompanied by differences in class-level F1 scores and confusion patterns, showing that aggregate accuracy alone does not fully describe model behaviour across the 30 scene categories.

Overall, the findings indicate that model selection involves a trade-off between predictive performance and computational cost. Under the experimental conditions of this study, Swin-T provided the highest observed mean classification performance with considerably lower parameter and FLOP requirements than ConvNeXt-Base, while MobileNetV3-L offered competitive accuracy with a substantially smaller model. These results support the use of multiple performance and efficiency measures when comparing aerial scene classification architectures.

The conclusions are limited to the dataset, training protocol, evaluated architectures, and computational setting considered in this study. Evaluation on additional datasets, geographic conditions, and hardware platforms would be required to assess whether the observed performance and efficiency patterns remain consistent in other settings.

REFERENCES

- [1] Zhu, X.X., Tuia, D., Mou, L.C., et al. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4): 8-36. <https://doi.org/10.1109/MGRS.2017.2762307>
- [2] Ma, L., Liu, Y., Zhang, X.L., Ye, Y.X., Yin, G.F., Johnson, B.A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152: 166-177. <https://doi.org/10.1016/j.isprsjprs.2019.04.015>
- [3] Jayanthi, S., Sathya, M.A.J., Balasubramanian, N., Karmakonda, K., Laxmikanthan, M., Manojkumar, V. (2026). Landslide susceptibility mapping in Idukki, Western Ghats, India by integrating weighted multi-criteria decision analysis and machine learning. *International Journal of Safety and Security Engineering*, 16(1): 65-76. <https://doi.org/10.18280/ijssse.160106>
- [4] Hu, F., Xia, G.S., Hu, J.W., Zhang, L.P. (2015). Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 7(11): 14680-14707. <https://doi.org/10.3390/rs71114680>
- [5] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778. <https://doi.org/10.1109/CVPR.2016.90>

- [6] Jayanthi, S., Sathya, M.A.J., Nathan, B., Karmakonda, K., Laxmikanthan, M., Manojkumar, V. (2026). Interpretable land-use classification using ResNet18 and SmoothGradCAM++: A study on the UC Merced dataset. *Ingénierie des Systèmes d'Information*, 31(2): 371-379. <https://doi.org/10.18280/isi.310205>
- [7] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2261-2269. <https://doi.org/10.1109/CVPR.2017.243>
- [8] Abouzahir, S., Sadik, M., Sabir, E. (2021). Bag-of-visual-words-augmented histogram of oriented gradients for efficient weed detection. *Biosystems Engineering*, 202: 179-194. <https://doi.org/10.1016/j.biosystemseng.2020.11.005>
- [9] Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2): 91-110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- [10] Tan, M.X., Le, Q.V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, PMLR, 97: 6105-6114. <https://proceedings.mlr.press/v97/tan19a.html>
- [11] Howard, A., Sandler, M., Chen, B., Wang, W.J., Chen, L.C., Tan, M.X. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 1314-1324. <https://doi.org/10.1109/ICCV.2019.00140>
- [12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. <https://doi.org/10.48550/arXiv.2010.11929>
- [13] Liu, Z., Lin, Y.T., Cao, Y., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9992-10002. <https://doi.org/10.1109/ICCV48922.2021.00986>
- [14] Bazi, Y., Bashmal, L., Rahhal, M.M.A., Dayil, R.A., Ajlan, N.A. (2021). Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3): 516. <https://doi.org/10.3390/rs13030516>
- [15] Maharana, K., Mondal, S., Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1): 91-99. <https://doi.org/10.1016/j.gltp.2022.04.020>
- [16] Zhou, B.L., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2921-2929. <https://doi.org/10.1109/CVPR.2016.319>
- [17] Liu, Z., Mao, H.Z., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.N. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 11966-11976. <https://doi.org/10.1109/cvpr52688.2022.01167>
- [18] Zhao, S.L., Gu, T.T., Wang, Z., Liu, X.W., Wang, Y.C., Feng, H.Q. (2026). Multi-scale remote sensing image classification algorithm based on self-attention and local attention. In *Proceedings of the 2026 International Symposium on Biological Neural Networks and Intelligent Optimization (BNNIO '26)*, ACM, New York, NY, USA, pp. 50-57. <https://doi.org/10.1145/3796551.3796559>
- [19] Aleissae, A.A., Kumar, A., Anwer, R.M., et al. (2023). Transformers in remote sensing: A survey. *Remote Sensing*, 15(7): 1860. <https://doi.org/10.3390/rs15071860>
- [20] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7: 1-30. <https://dl.acm.org/doi/10.5555/1248547.1248548>
- [21] Bouthillier, X., Delaunay, P., Bronzi, M., et al. (2021). Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 3: 747-769. <https://arxiv.org/abs/2103.03098>
- [22] Xia, G.S., Hu, J., Hu, F., et al. (2017). AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7): 3965-3981. <https://doi.org/10.1109/TGRS.2017.2685945>
- [23] Kaggle. AID scene classification datasets. <https://www.kaggle.com/datasets/jiayuanhengala/aid-scene-classification-datasets>
- [24] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7132-7141. <https://doi.org/10.1109/CVPR.2018.00745>
- [25] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2): 153-157. <https://doi.org/10.1007/BF02295996>
- [26] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6): 80-83. <https://doi.org/10.2307/3001968>
- [27] McInnes, L., Healy, J., Saul, N., Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29): 861. <https://doi.org/10.21105/joss.00861>