**JURNAL RISET INFORMATIKA** **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 8, No. 3. Juni 2026 DOI: https://doi.org/10.34288/jri.v8i3.515
**Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023**

# TRANSFER LEARNING ARCHITECTURE SELECTION FOR REMOTE SENSING SCENE CLASSIFICATION

**Akhiyar Waladi[-1], Hasanatul Iftitah[-2]**

Department of Informatics, Faculty of Science and Technology
Universitas Jambi
akhiyar.waladi@unja.ac.id [1], hasanatul.iftitah@unja.ac.id [2]

**Abstract**

Selecting a deep learning architecture for classifying remote sensing scenes usually involves comparing published accuracy across papers that each use different training protocols, making it unclear whether accuracy gaps reflect architecture or training differences. This study aims to examine whether such gaps remain meaningful when architectures are evaluated under one uniform training protocol. We isolate the architecture variable by evaluating eight models from three design families, five classical CNNs (ResNet-50, ResNet-101, DenseNet-121, EfficientNet-B0, EfficientNet-B3), two vision transformers (ViT-B/16, Swin Transformer), and one modernized CNN (ConvNeXt-Tiny), under identical training conditions on EuroSAT (10 classes, 27,000 Sentinel-2 patches) and UC Merced (21 classes, 2,100 aerial photographs). Every model shares the same ImageNet-1K initialization, AdamW optimizer, augmentation pipeline, and early stopping rule. ConvNeXt-Tiny reached the highest accuracy on EuroSAT (99.11%) and Swin-T on UC Merced (99.76%), but the accuracy range was only 0.41 percentage points (1.66 on UC Merced). McNemar's test confirmed that most pairwise differences were not significant. EfficientNet-B0, the smallest model at 4.0M parameters, reached 98.76% and 99.52% while using 21x fewer parameters than ViT-B/16. On these two well-studied benchmarks, a single uniform training configuration was sufficient to bring all architectures to near-identical performance. This convergence suggests that, under the selected protocol and on these saturated benchmarks, architecture choice had a limited effect on classification performance compared with the influence of the shared training procedure. Further studies are needed to determine whether this convergence persists on more challenging benchmarks, architecture-specific optimal configurations, or domain-specific pretraining schemes.

Keywords: Remote Sensing; Scene Classification; Transfer Learning; Vision Transformer; Benchmark Comparison

***Abstrak***

*Pemilihan arsitektur deep learning untuk klasifikasi scene penginderaan jauh umumnya dilakukan dengan membandingkan akurasi yang dipublikasikan dari berbagai paper yang masing-masing menggunakan protokol pelatihan berbeda, sehingga sulit dipastikan apakah perbedaan akurasi mencerminkan arsitektur atau prosedur pelatihan. Penelitian ini bertujuan untuk menguji apakah perbedaan tersebut masih bermakna ketika berbagai arsitektur dievaluasi menggunakan satu protokol pelatihan yang seragam. Penelitian ini mengisolasi variabel arsitektur dengan mengevaluasi delapan model dari tiga keluarga desain, yaitu lima CNN klasik (ResNet-50, ResNet-101, DenseNet-121, EfficientNet-B0, EfficientNet-B3), dua vision transformer (ViT-B/16, Swin Transformer), dan satu CNN modern (ConvNeXt-Tiny), dalam kondisi pelatihan identik pada dataset EuroSAT (10 kelas, 27.000 citra Sentinel-2) dan UC Merced (21 kelas, 2.100 foto udara). Setiap model menggunakan inisialisasi ImageNet-1K, optimizer AdamW, pipeline augmentasi, dan aturan early stopping yang sama. ConvNeXt-Tiny mencapai akurasi tertinggi pada EuroSAT (99,11%) dan Swin-T pada UC Merced (99,76%), namun rentang akurasi hanya 0,41 poin persentase pada EuroSAT dan 1,66 pada UC Merced. Uji McNemar mengonfirmasi bahwa sebagian besar perbedaan berpasangan tidak signifikan. EfficientNet-B0, model terkecil dengan 4,0 juta parameter, mencapai 98,76% dan 99,52% dengan menggunakan 21x lebih sedikit parameter dibandingkan ViT-B/16. Pada dua benchmark yang banyak diteliti ini, satu resep pelatihan seragam cukup membawa semua arsitektur ke kinerja yang hampir identik. Konvergensi ini menunjukkan bahwa pada tugas klasifikasi yang telah jenuh, dalam protokol dan benchmark yang digunakan, pemilihan arsitektur memiliki pengaruh terbatas. Apakah konvergensi yang sama berlaku pada benchmark yang lebih sulit, konfigurasi optimal khusus arsitektur, atau pretraining spesifik domain masih perlu diuji.*

*Kata kunci: Penginderaan Jauh; Klasifikasi Scene; Transfer Learning; Vision Transformer; Perbandingan Benchmark*

## INTRODUCTION

Scene classification, the task of assigning one land-use label to an entire satellite or aerial patch, underpins applications from urban growth monitoring to disaster response (Cheng et al., 2020; Ma et al., 2019). Practitioners routinely need to pick a model, but published accuracy numbers come from papers that each tune their own training protocol, so the comparisons are hard to trust.

The field went through two shifts. First, CNNs replaced the older bag-of-visual-words and SVM pipelines (Yang & Newsam, 2010) that required manual feature design and broke down beyond a few dozen classes (Cheng et al., 2017). ResNet (He et al., 2016), DenseNet (Huang et al., 2017), and their successors learn features directly from pixels. Initializing these networks with ImageNet (Deng et al., 2009) pretraining weights consistently outperforms training from scratch (Li et al., 2018; Nogueira et al., 2017), largely because labeled remote sensing data is small relative to what the models need (Neumann et al., 2019; Pan & Yang, 2010).

Vision transformers (Dosovitskiy et al., 2021; Liu et al., 2021) have been tested on remote sensing classification, but their advantage varies across studies (Aleissaee et al., 2023; Bazi et al., 2021; Hong et al., 2022; Wang et al., 2023). Purely convolutional alternatives like ConvNeXt (Liu et al., 2022) have matched transformer accuracy on ImageNet without using attention layer. Wightman et al. (Wightman et al., 2021) further showed that a standard ResNet-50 gained 4.3 percentage points on ImageNet by adopting a modern training schedule alone, without architectural modification. These developments raise the question of how much the architecture itself contributes when the training procedure is held constant.

Architectures in this field belong to three families (classical CNNs, vision transformers, and modernized CNNs), and each has papers claiming superiority. But most comparisons test only two or three models, use different training protocols, or evaluate on a single dataset (Adegun et al., 2023). Large-scale studies like Battle of the Backbones (Goldblum et al., 2023) covered general computer vision tasks but did not focus on remote sensing under controlled conditions.

We designed an experiment to isolate the architecture variable on two remote sensing benchmarks. All eight architectures, covering the three families, are trained with the same ImageNet-1K initialization, augmentation, AdamW optimizer, and early stopping rule. We test on EuroSAT (medium-resolution Sentinel-2 optical imagery, 10 classes) and UC Merced (high-resolution aerial photographs, 21 classes), then apply McNemar's test (Dietterich, 1998; McNemar, 1947) to every pairwise comparison.

This paper compares eight architectures from three design families on two remote sensing benchmarks using identical training conditions. By fixing the optimizer, initialization, augmentation, and early stopping rule, the study aims to isolate architecture as the main experimental variable. The comparison is designed to examine whether performance differences results among modern architectures remain meaningful on saturated benchmarks such as EuroSAT and UC Merced. In addition to reporting overall accuracy, F1-Macro, and Cohen's kappa, this study applies McNemar's test to all 28 model pairs per dataset and analyzes parameter efficiency by comparing model size and training time. This design provides a controlled basis for further interpreting architecture selection without presenting the background section as a conclusion of the experimental results.

We note upfront that the experiment uses a single train/test split, a fixed CNN-style training configuration, and only ImageNet-1K pretraining, which limits the generalizability of the conclusions. These constraints are further explained in the Discussion section.

## RELATED WORK

### A. CNN-Based Scene Classification

Residual connections, proposed by He et al. (He et al., 2016), enabled stable training beyond 100 layers. ResNet-50 remains widely adopted for remote sensing classification tasks (Adegun et al., 2023). Huang et al. (Huang et al., 2017) created dense connectivity that maximizes feature reuse with a small parameter budget. Tan and Le (Tan & Le, 2019) showed that scaling depth, width, and resolution together outperforms scaling any one dimension alone; EfficientNet-B0 achieved the best accuracy-to-compute ratio on ImageNet at the time.

Six architectures have been compared across three aerial datasets. Nogueira et al. (Nogueira et al., 2017) tested three transfer strategies (full training, fixed feature extraction, and end-to-end fine-tuning) and reported that fine-

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.515
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

tuning pretrained weights outperformed training from random initialization in every case (Ma et al., 2019; Zhu et al., 2017). Wightman et al. (Wightman et al., 2021) retrained a vanilla ResNet-50 with a modernized configuration and raised its ImageNet accuracy from 76.1% to 80.4% without changing a single layer. This result implies that many reported architecture comparisons are confounded by training protocol differences.

**B. Transformer-Based Approaches**

The Vision Transformer (ViT) divides each image into 16x16 non-overlapping patches and processes the resulting sequence through multi-head self-attention (Dosovitskiy et al., 2021). To reduce the quadratic computational memory cost, Swin Transformer restricts attention to small local windows that shift between layers, bringing complexity down to linear (Liu et al., 2021). Both architectures have been adopted in remote sensing: Bazi et al. (Bazi et al., 2021) evaluated ViT for scene classification, Hong et al. (Hong et al., 2022) adapted the transformer paradigm to hyperspectral imagery, and a survey of over 60 transformer-based methods by Aleissaee et al. (Aleissaee et al., 2023) concluded that standalone transformers generally require large training sets to outperform CNNs.

**C. Modernized CNNs**

Liu et al. (Liu et al., 2022) started from a plain ResNet and adopted one transformer design choice at a time: patchify stem, 7x7 depthwise kernels, LayerNorm with GELU, and inverted bottleneck. Self-attention was never added, yet the final model reached similar performance to Swin Transformer on ImageNet and COCO.

**D. Benchmark Datasets**

UC Merced (Yang & Newsam, 2010) contains 2,100 aerial photographs at 0.3 m resolution split across 21 land-use classes. EuroSAT (Helber et al., 2019) is much larger (27,000 Sentinel-2 patches, 10 m, 10 classes) but coarser in spatial detail. Both are small enough to run dozens of experiments on a single GPU, which is why they remain popular despite the existence of bigger collections like NWPU-RESISC45 (Cheng et al., 2017) (31,500 images across 45 categories) and AID (10,000 images across 30 categories) (Xia et al., 2017). The downside is saturation. Cheng et al. already warned in 2017 that UC Merced accuracies had plateaued. EuroSAT appears to be following a similar trajectory, with multiple recent studies reporting transfer learning accuracies above 98%.

We chose these two datasets precisely because of this plateau. Our reasoning was that if eight architectures all land above 98% on a pair of benchmarks whose imaging modalities and class granularities differ substantially (0.3 m aerial with 21 classes versus 10 m satellite with 10 classes), the remaining accuracy differences may not necessarily indicate substantial architectural advantages. Confirming or refuting this hypothesis on datasets where the performance ceiling is lower, such as NWPU-RESISC45, is left to future work.

**E. Self-Supervised Pretraining**

A growing body of work has begun to replace standard ImageNet-supervised weights with representations learned directly from satellite and aerial imagery through self-supervised objectives. Methods such as SeCo (Neumann et al., 2019) pretrain on large volumes of unlabeled remote sensing data and produce features that are better aligned with the spectral and spatial statistics of Earth observation imagery than features learned from natural photographs. Because ViT was originally conceived for self-supervised pretraining at scale, domain-specific weights could shift the relative ranking of architectures in ways that an ImageNet-1K starting point does not capture. None of these pretraining alternatives were tested in the present work, and all results reported here should be read in the context of ImageNet-1K supervised transfer learning only.

**F. Statistical Testing**

Dietterich (Dietterich, 1998) compared five statistical tests for classifier comparison and found that McNemar's test has acceptable Type I error when classifiers run once on a fixed split. Foody (Foody, 2004) applied it to remote sensing map comparison. Despite these recommendations, most deep learning papers in remote sensing compare models by accuracy tables alone (Adegun et al., 2023).

## RESEARCH METHODS

The experimental pipeline is shown in Figure 1. Both datasets go through the same preprocessing, all eight architectures are trained with identical settings, and the resulting models are compared on accuracy, statistical significance, error patterns, and training cost.

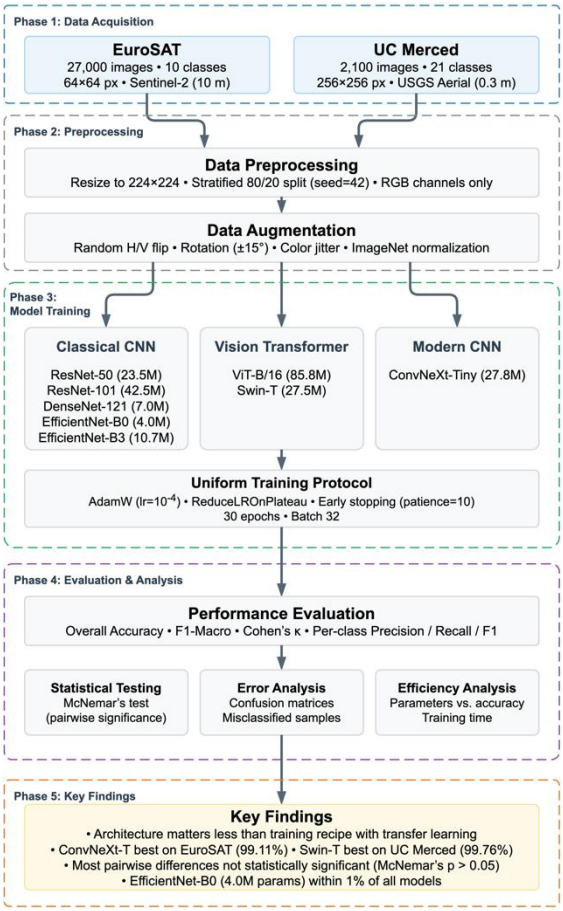


Figure 1. Overview of the research methodology.

## A. Datasets

### 1) EuroSAT

EuroSAT dataset (Helber et al., 2019) contains 27,000 geo-referenced image patches from Sentinel-2 imagery, organized into 10 land-use categories (see Figure 2). Each patch is 64x64 pixels at 10 m ground sampling distance.

We use only the RGB bands (B4, B3, B2) because all eight architectures expect three-channel input. This choice removes 10 of the 13 available spectral bands, reducing relevance to multispectral workflows while preserving fair comparison across architectures built for three-channel input.

### 2) UC Merced Land Use

UC Merced (Yang & Newsam, 2010) consists of 2,100 overhead photographs acquired from the USGS National Map at 0.3 m ground resolution and organized into 21 land-use categories with 100 images each (256x256 pixels). Figure 3. Using this spatial resolution, individual houses, cars, plane and tennis courts are clearly discernible, yet sub-categories (denseresidential, mediumresidential, sparseresidential) which is residential still remain difficult to separate because they contain the same building types and differ mainly in the spacing between structures.


Figure 2. EuroSAT dataset samples. One Sentinel-2 patch per class (64x64, 10 m, 10 classes)

### 3) Data Partitioning and Validation

We applied an 80/20 stratified random split with seed 42, producing 21,600 training and 5,400 test images for EuroSAT, and 1,680 training and 420 test images for UC Merced. The test set also serves as the validation set for triggering early stopping and adjusting the learning rate. This is a common best practice in scene classification task

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.515
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

benchmarks but constitutes an optimistic bias, as the model selection criterion is evaluated on the same data used for final reporting.


Figure 3. UC Merced dataset samples. One aerial image per class (256x256, 0.3 m, 21 classes)

## B. Model Architectures

We selected eight architectures covering three design families (Table 1). The selection includes in total, five classical CNNs, two vision transformers, and one modernized CNN. The larger number of CNN-based models reflects their frequent use as standard baselines in remote sensing scene classification, while ViT-B/16, Swin-T, and ConvNeXt-Tiny were included to represent transformer-based and modernized CNN designs. We note that CNNs are overrepresented (six models if ConvNeXt is counted as a CNN), which limits the balance of the family-level comparison.

Table 1. Model architectures

| Model | Family | Param | Year |
|---|---|---|---|
| ResNet-50 | CNN | 23,5 | 2016 |
| ResNet-101 | CNN | 42,5 | 2016 |
| DenseNet-121 | CNN | 7 | 2017 |
| EfficientNet-B0 | CNN | 4 | 2019 |
| EfficientNet-B3 | CNN | 10,7 | 2019 |
| ViT-B/16 | Transformer | 85,8 | 2021 |
| Swin-T | Transformer | 27,5 | 2021 |
| ConvNeXt-T | Modern CNN | 27,8 | 2022 |

**Classical CNNs**. ResNet-50 and ResNet-101 use residual connections. DenseNet-121 connects each layer to every preceding layer. EfficientNet-B0 and B3 apply compound scaling with depthwise separable convolutions.

**Vision Transformers**. ViT-B/16 splits each 224x224 image into 196 tokens and processes them through 12 self-attention layers (85.8M parameters). Swin-T restricts attention to local windows that shift between layers (27.5M parameters).

**Modernized CNN**. ConvNeXt-Tiny uses 7x7 depthwise kernels, LayerNorm, GELU, and an inverted bottleneck layout. These design elements were adopted from the transformer literature but are implemented entirely with convolutions (27.8M parameters).

All model training starts from ImageNet-1K pretrained weights (IMAGENET1K_V1) loaded through torchvision.models (Paszke et al., 2019). We replace the final classification head with a linear layer matching the target class count.

## C. Training Protocol

Every model trains with the same settings so that any accuracy difference comes from the architecture, not the optimizer or augmentation. Each image undergoes bilinear interpolation to reach 224x224 pixels (upsampling 3.5x for EuroSAT's 64x64 patches). Augmentation includes horizontal and vertical flips applied at random, image rotation up to range of +/-15 degrees, and photometric perturbation (brightness and contrast +/-0.2, saturation +/-0.1). Each channel is then normalized to ImageNet mean and standard deviation. Figure 4 shows augmentation pipeline applied to EuroSAT samples: the leftmost column contains original images resized to 224x224 pixels. The other columns present the corresponding

randomly augmented images. These geometric and photometric perturbations expand the effective training set and reduce overfitting, which is particularly important for UC Merced where each class has only 80 training images.

We optimize with AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of $10^{-4}$ and weight decay of $10^{-4}$. The learning rate halves when test accuracy stalls for five epochs, and training stops after ten epochs of no improvement. Batch size is 32, maximum epochs is 30, and all runs use a single NVIDIA RTX GPU with PyTorch 2.0.

This protocol is intentionally uniform, but we acknowledge it may not be equally suited to all architectures. The augmentation is moderate by current standards: ViT was originally trained with MixUp, CutMix, RandAugment, and larger batch sizes (Dosovitskiy et al., 2021), while ConvNeXt used similar strong augmentation (Liu et al., 2022). Our training protocol is closer to a standard CNN configuration. Furthermore, the implications of these settings are explained in the Discussion section.

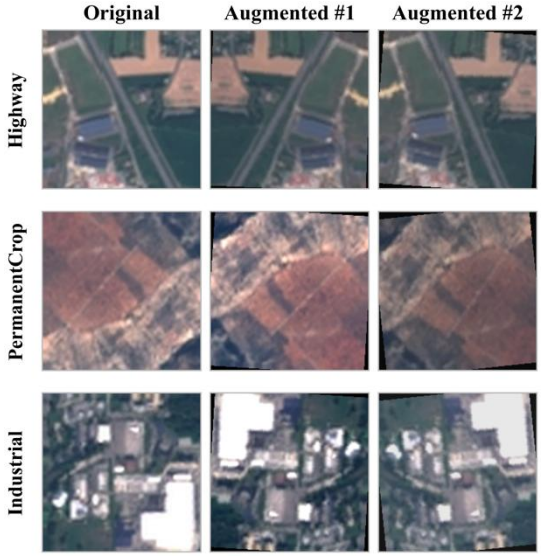


Figure 4. Data augmentation pipeline

### D. Evaluation Metrics

#### 1) Classification Metrics

We report overall accuracy (OA), macro-averaged F1-score (F1-Macro, unweighted mean of per-class F1 values), and Cohen's kappa (kappa) (Cohen, 1960), which adjusts for chance agreement (Congalton, 1991; Foody, 2002).

#### 2) Statistical Significance

We use McNemar's test (McNemar, 1947) with continuity correction, following Dietterich's (Dietterich, 1998) recommendation and Foody's (Foody, 2004) application to remote sensing:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \quad (1)$$

where $n_{01}$ counts samples correctly classified by model A but missed by B, and $n_{10}$ the reverse. We use significance thresholds of $\alpha$ = 0.05 and 0.01.

#### 3) Computational Efficiency

For each model and dataset, we record the trainable parameter count and total wall-clock training time. The recorded time covers all epochs up to the point where early stopping triggered, so it reflects actual training cost rather than a fixed epoch budget.

## RESULTS AND DISCUSSION

### A. Overall Performance

Table 2 and Table 3 list overall accuracy (OA), F1-Macro, and Cohen's kappa for both datasets. On EuroSAT, ConvNeXt-Tiny leads at 99.11% OA (kappa = 0.990), followed by Swin-T (99.02%) and EfficientNet-B3 (98.98%). ResNet-50 sits at the bottom with 98.70%. The gap between the best and worst model is just 0.41 percentage points, and F1-Macro mirrors this tight spread (0.9908 versus 0.9865). On UC Merced, Swin-T leads at 99.76% with four models tied at 99.52%. ViT-B/16 trails at 98.10%, widening the range to 1.66 points. Kappa values exceed 0.98, confirming near-perfect agreement on both datasets.

Table 2. Performance EuroSAT (5,400 test images)

| Model | OA (%) | F1-Macro | κ | M |
|---|---|---|---|---|
| **ConvNeXt-T** | **99,11** | **0,990** | 0,990 | 27.8 |
| Swin-T | 99,02 | 0,989 | 0,989 | 27.5 |
| EfficientNet-B3 | 98,98 | 0,989 | 0,988 | 10.7 |
| DenseNet-121 | 98,96 | 0,989 | 0,988 | 7.0 |
| ResNet-101 | 98,91 | 0,988 | 0,987 | 42.5 |
| EfficientNet-B0 | 98,76 | 0,987 | 0,986 | 4.0 |
| ViT-B/16 | 98,72 | 0,987 | 0,985 | 85.8 |
| ResNet-50 | 98,7 | 0,986 | 0,985 | 23.5 |

No model fell below 98% on either dataset. In absolute numbers, ConvNeXt-Tiny misclassified 48 of 5,400 EuroSAT test patches whereas ResNet-50 misclassified 70, a difference of just 22 images. On UC Merced the margin is even thinner: Swin-T made a single error out of 420 test images, while ViT-B/16 made eight. These figures are broadly consistent with published transfer learning results,

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.515
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

which typically fall between 97% and 99% on EuroSAT (Helber et al., 2019; Neumann et al., 2019) and above 97% on UC Merced (Cheng et al., 2017). Figure 5 narrow distribution is clearly visible.

Table 3. Performance UC Merced (420 test images)

| Model | OA (%) | F1-Macro | κ | M |
|---|---|---|---|---|
| **Swin-T** | **99,76** | **0,997** | 0,997 | 27.5 |
| ConvNeXt-T | 99,52 | 0,995 | 0,995 | 27.8 |
| EfficientNet-B0 | 99,52 | 0,995 | 0,995 | 4.0 |
| EfficientNet-B3 | 99,52 | 0,995 | 0,995 | 10 |
| ResNet-50 | 99,52 | 0,995 | 0,995 | 23.5 |
| ResNet-101 | 99,29 | 0,992 | 0,992 | 42.5 |
| DenseNet-121 | 99,05 | 0,990 | 0,99 | 7.0 |
| ViT-B/16 | 98,1 | 0,980 | 0,98 | 85.8 |

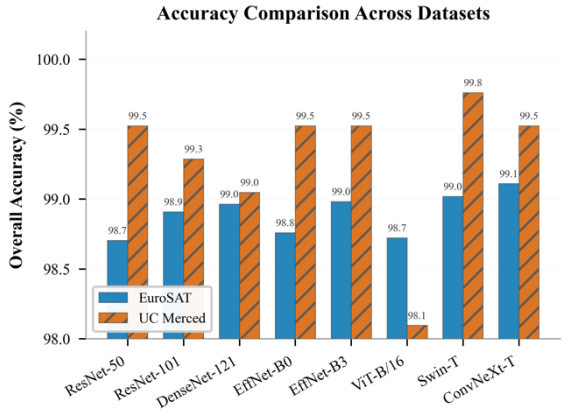


Figure 5. Accuracy comparison both datasets

## B. Training Dynamics

Figure 6 plots the validation loss and accuracy curves for EuroSAT. All eight models follow a similar downward loss trajectory with epoch-to-epoch oscillations. Only two models trigger early stopping: Swin-T reaches its best accuracy at epoch 13 and stops at epoch 23, while ResNet-50 peaks at epoch 19 and stops at epoch 29. The remaining six models train for the full 30 epochs without triggering the patience threshold.

On UC Merced (Figure 7), convergence is faster due to the smaller training set (1,680 images). EfficientNet-B3 starts with a validation loss above 2.0 but drops sharply by epoch 5. DenseNet-121 and ViT-B/16 both reach their best accuracy at epoch 6, with early stopping firing at epoch 16. With 85.8M parameters fine-tuned on just 1,680 images, ViT-B/16 did not have enough training iterations to adapt its pretrained attention patterns to 0.3 m aerial imagery. Swin-T and ResNet-50 are the only models that train for all 30 epochs, and Swin-T achieves the best final accuracy (best epoch 27).

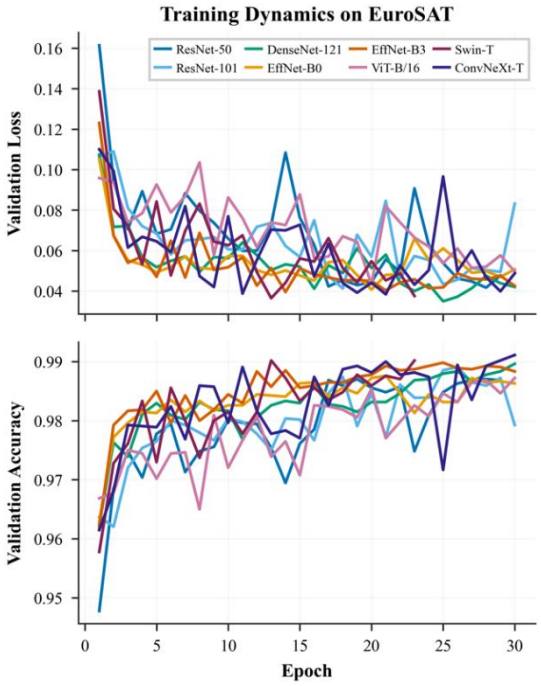


Figure 6. Training dynamics on EuroSAT

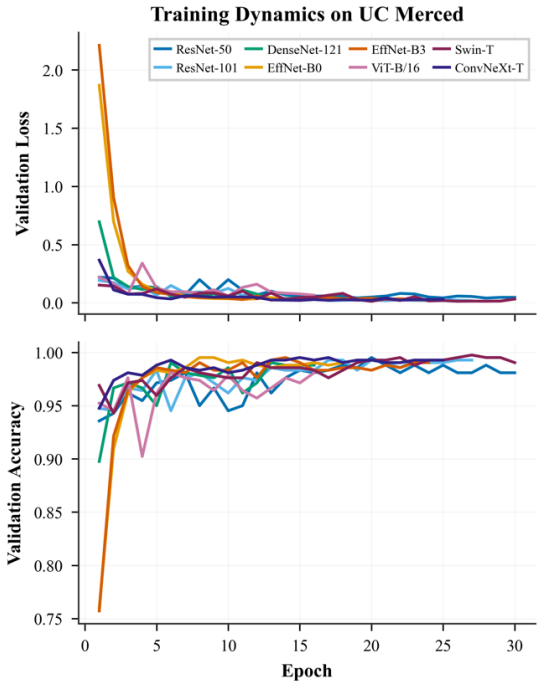


Figure 7. Training dynamics on UC Merced

## C. Per-Class Analysis

Table 4 and Table 5 report per-class F1-scores for all eight models. On EuroSAT (Table 4), SeaLake and Residential are classified near-

perfectly (F1 >= 0.99) by every architecture, while PermanentCrop and Pasture show the widest spread. PermanentCrop F1 ranges from 0.97 (ResNet-50, EfficientNet-B0, ViT-B/16) to 0.98 (Swin-T, ConvNeXt-T), a gap that is small in absolute terms but consistent across repeated class confusions with HerbaceousVegetation, as report in the discussion section. River patches also exhibit above-average variance, most likely because thin watercourses occupying part of a 64x64 pixel tile can be spectrally and texturally similar to highway segments.

On UC Merced (Table 5), five classes (agricultural, beach, chaparral, harbor, parkinglot) achieved F1 = 1.00 across all models and are omitted. Among the remaining 16 classes, ViT-B/16 shows the largest deviations: buildings (0.92), denseresidential (0.92), mediumresidential (0.93), and storagetanks (0.95). At 0.3 m resolution the three residential sub-categories contain visually similar rooftop and road patterns and differ primarily in building density, a distinction that proves difficult even for human interpreters in borderline cases.

### D. Statistical Significance

The McNemar p-value matrix for the EuroSAT test set is shown in Figure 8. Of the 28 unique model pairs, the majority yielded $p > 0.05$ (dark purple cells), indicating that the observed accuracy differences are within the range of sampling variability for this particular test partition.

Table 4. Per-class F1-scores on EuroSAT (5,400 test images)

| Class | RN-50 | RN-101 | DN-121 | EN-B0 | EN-B3 | ViT | Swin-T | CNX-T |
|---|---|---|---|---|---|---|---|---|
| AnnualCrop | .98 | .98 | .98 | .98 | **.99** | .98 | .98 | .98 |
| Forest | .99 | **1.0** | **1.0** | .99 | **1.0** | .99 | **1.0** | **1.0** |
| HerbVeg | .98 | .98 | .98 | **.99** | .98 | .98 | .98 | **.99** |
| Highway | .98 | **.99** | **.99** | **.99** | **.99** | **.99** | **.99** | **.99** |
| Industrial | .99 | .99 | **1.0** | .99 | .99 | .99 | **1.0** | .99 |
| Pasture | .98 | .98 | .98 | .97 | .98 | .98 | .98 | **.99** |
| PermCrop | .97 | .97 | .97 | .97 | **.98** | .97 | **.98** | **.98** |
| Residential | **1.0** | **1.0** | **1.0** | .99 | 1.0 | 1.0 | 1.0 | 1.0 |
| River | .99 | .99 | .99 | .99 | .99 | .99 | **1.0** | .99 |
| SeaLake | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |

Table 5. Per-class F1-scores on UC Merced (420 test images, 21 classes shown)

| Class | RN-50 | RN-101 | DN-121 | EN-B0 | EN-B3 | ViT | Swin-T | CNX-T |
|---|---|---|---|---|---|---|---|---|
| agricultural | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| airplane | **1.0** | .97 | .97 | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| baseballdmnd | .98 | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| beach | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| buildings | **1.0** | **1.0** | **1.0** | .95 | .98 | .92 | **1.0** | **1.0** |
| chaparral | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| denseRes | .97 | .97 | .97 | .98 | .98 | .92 | **1.0** | .98 |
| forest | **1.0** | **1.0** | .98 | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| freeway | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | .97 | .97 |
| golfcourse | .97 | .97 | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| harbor | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| intersection | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | .98 | **1.0** | **1.0** |
| mediumRes | .98 | .98 | .95 | **1.0** | .97 | .93 | **1.0** | **1.0** |
| mobileHome | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | .97 | **1.0** | .97 |
| overpass | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | .97 | .98 | .98 |
| parkinglot | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| river | **1.0** | .98 | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| runway | **1.0** | .98 | .98 | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** |
| sparseRes | **1.0** | **1.0** | .95 | .97 | .97 | .97 | **1.0** | **1.0** |
| storagetanks | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | .95 | **1.0** | **1.0** |
| tenniscourt | **1.0** | **1.0** | **1.0** | **1.0** | **1.0** | .98 | **1.0** | **1.0** |

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.515
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Significant differences (orange cells) involved only four of the 28 pairs: ConvNeXt-Tiny versus ResNet-50 ($p < 0.01$) and ViT-B/16 ($p < 0.01$), ConvNeXt-Tiny versus EfficientNet-B0 ($p < 0.05$), and Swin-T versus ResNet-50 ($p < 0.05$). The remaining 24 pairs did not reach significance at the unadjusted alpha level, suggesting that most observed accuracy gaps were small under this test partition. However, given that each dataset involved 28 pairwise comparisons, these p-values should be interpreted carefully. Because this study used no multiple-comparison correction, such as Bonferroni or Holm correction, the limited significant outcomes should be understood as exploratory evidence rather than confirmatory conclusions.

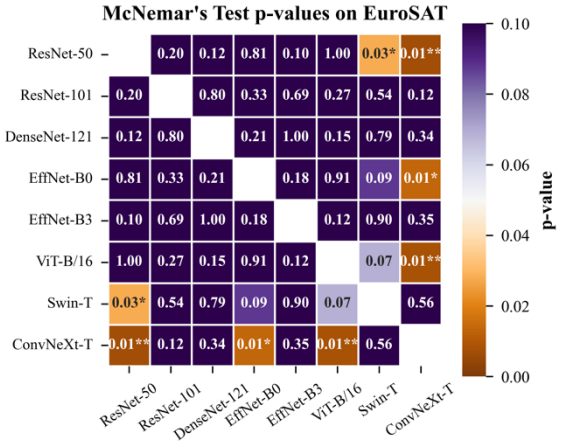


Figure 8. McNemar's test p-value matrix for EuroSAT. Purple cells indicate low p-values

On UC Merced, only a single pair reached significance: Swin-T versus ViT-B/16 ($p = 0.05$). The remaining 27 pairs were non-significant despite the wider accuracy spread (1.66 points), a consequence of the limited test set size (420 images, 20 per class). With so few samples, the McNemar test lacks statistical power to detect small but potentially real differences between models. This Type II error risk should be kept in mind when interpreting the non-significant outcomes (see the Discussion section).

### E. Error Analysis

The most frequent confusion pairs across all models were PermanentCrop misclassified as HerbaceousVegetation (63 times), AnnualCrop as PermanentCrop (55), and Pasture as AnnualCrop (32). River confused with Highway added another 27 errors. These classes share similar spectral profiles at 10 m resolution. On UC Merced, errors were far rarer (1-8 per model) and concentrated among the residential sub-types (5 errors), freeway versus overpass, and airplane versus runway. These EuroSAT errors clustered around the four vegetation classes and were consistent with the per-class F1 scores in Table 5.

Figure 9 and Figure 10 place misclassified test images (red border) alongside correctly classified examples from the predicted class (blue border). In the EuroSAT pairs, the misclassified PermanentCrop patches contain mixed crop rows and vegetated margins that are often visually indistinguishable from HerbaceousVegetation at 64x64 pixels. The River-Highway pair shows that linear water features can resemble road segments at this resolution. On UC Merced, denseresidential and mediumresidential patches share the same rooftop geometry and road layouts, differing only in building density. Since all eight architectures produced the same confusion patterns, these errors are better attributed to inherent class overlap in the datasets than to shortcomings in any one model.

Table 6. Computational efficiency comparison. Models sorted by parameter count

| Model | Family | Parameters (Million) | OA (%) | | Time (s) | |
|---|---|---|---|---|---|---|
| | | | Euro | UCM | Euro | UCM |
| **EfficientNet-B0** | CNN | **4** | 98,76 | 99,52 | **1.088** | 216 |
| DenseNet-121 | CNN | 7 | 98,96 | 99,05 | 1.686 | **211** |
| EfficientNet-B3 | CNN | 10,7 | 98,98 | 99,52 | 1.559 | 312 |
| ResNet-50 | CNN | 23,5 | 98,7 | 99,52 | 1.630 | 402 |
| Swin-T | Trans | 27,5 | **99,02** | **99,76** | 1.474 | 421 |
| ConvNeXt-T | Mod | 27,8 | 99,11 | 99,52 | 2.470 | 385 |
| ResNet-101 | CNN | 42,5 | 98,91 | 99,29 | 1.928 | 374 |
| ViT-B/16 | Trans | 85,8 | 98,72 | 98,1 | 3403 | 282 |

**F. Computational Efficiency**

Table 6 show a combination of parameter counts, accuracy on both datasets, and wall-clock training times. EfficientNet-B0 stands out as the smallest model, with 4.0M parameters with 98.76% on EuroSAT and 99.52% on UC Merced. ViT-B/16, despite having 85.8M parameters (21x more), scored 98.72% on EuroSAT and 98.10% on UC Merced. In training speed, EfficientNet-B0 was the fastest on EuroSAT (1,088 s) while ViT-B/16 was the slowest (3,403 s). On UC Merced, ViT-B/16 finished in only 282 s because early stopping fired at epoch 6. The Training times are not directly comparable across models because early stopping terminates each model at a different epoch; per-epoch cost would be a fairer measure, which we did not record in this study.

## DISCUSSION

**A. Architecture on Saturated Benchmarks**

ConvNeXt-Tiny, a modernized CNN model, achieved the highest accuracy on EuroSAT dataset (99.11%). Swin-T, a hierarchical transformer, ranked first on UC Merced (99.76%). Classical CNNs were distributed across both the middle and lower tiers in each ranking. The lack of a consistent winner across the two datasets suggests that, at this accuracy level and under this particular training protocol, the choice of architecture family is not a decisive factor.

Comparing the two of transformer-based models reveals a particularly informative contrast. Swin-T ranked second on EuroSAT and first on UC Merced, while ViT-B/16 ranked seventh and eighth, respectively. The two models differ in how they process local versus global image regions. Swin-T restricts attention computation to local, non-overlapping blocks of the feature map rather than processing the full image at once. This process builds a four-level pyramid at progressively lower resolutions, mirroring the multi-scale output of conventional convolutional encoders. ViT-B/16, by contrast, computes global self-attention across all 196 input tokens without any multi-scale hierarchy. In scene classification tasks where each 224x224 patch depicts a single land-use category with relatively uniform texture, global receptive fields appear to offer little benefit over local feature extraction.

Global self-attention shows its limits on small fine-tuning sets, as reflected by ViT-B/16's 98.10% on UC Merced. The model contains 85.8M trainable parameters but was fine-tuned on only 1,680 images (80 per class), and early stopping terminated training after just 6 epochs. This outcome is consistent with observations by (Dosovitskiy et al., 2021) et al. , who reported that ViT models require substantially larger pretraining corpora to surpass CNN performance. Neyshabur et al. (Neyshabur et al., 2020) further demonstrated that fine-tuned weights tend to remain in the vicinity of the pretrained initialization. One possible explanation is that ViT-B/16 remained strongly influenced by its ImageNet-1K pretrained model representations, which may have constrained its adaptation to 0.3 m aerial imagery. However, this explanation remains tentative because attention patterns and learned feature representations were not directly examined in this study.

**B. Training Configuration Fairness**

Using one training protocol for all eight models keeps the comparison controlled, but it may disadvantage architectures originally developed with different training settings. The protocol adopted in this study (flips, mild rotation, color jitter, AdamW at $10^{-4}$, batch size 32, without MixUp, CutMix, or label smoothing) reflects a conventional CNN fine-tuning configuration. ViT-B/16, however, was originally developed under substantially heavier regularization with batch sizes exceeding 4,000 (Dosovitskiy et al., 2021), and ConvNeXt similarly benefited from aggressive data augmentation during its initial ImageNet training (Liu et al., 2022). Given these differences, ViT-B/16 likely did not reach its full potential under our protocol, which may partly explain its lower ranking compared to the CNN-family models.

The final performance of ViT-B/16 might improve if it were trained with a configuration closer to the standard ViT recipe. In the original ViT setting, stronger regularization and augmentation strategies, such as MixUp, CutMix, label smoothing, repeated augmentation, and larger batch sizes, are commonly used to stabilize training and reduce overfitting. Such a configuration could help ViT-B/16 adapt its global attention mechanism more effectively to remote sensing images, especially on small datasets such as UC Merced. Therefore, the lower ranking of ViT-B/16 in this study should not be interpreted as evidence that ViT is inherently unsuitable for remote sensing scene classification. Rather, it indicates that ViT may be more sensitive to training configuration setup than CNN-based architectures under the uniform protocol used.

Our data only address the CNN side of this question, where the five convolutional models converged to nearly the same accuracy under the shared protocol.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.515
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

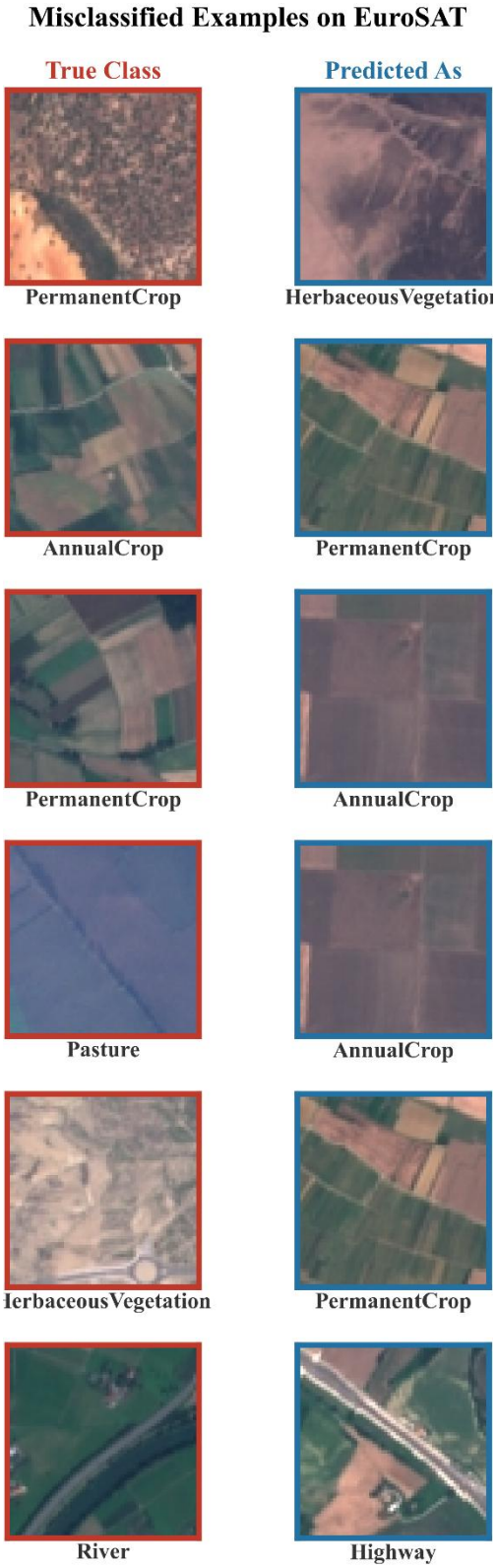


Figure 9. Misclassified examples from EuroSAT.

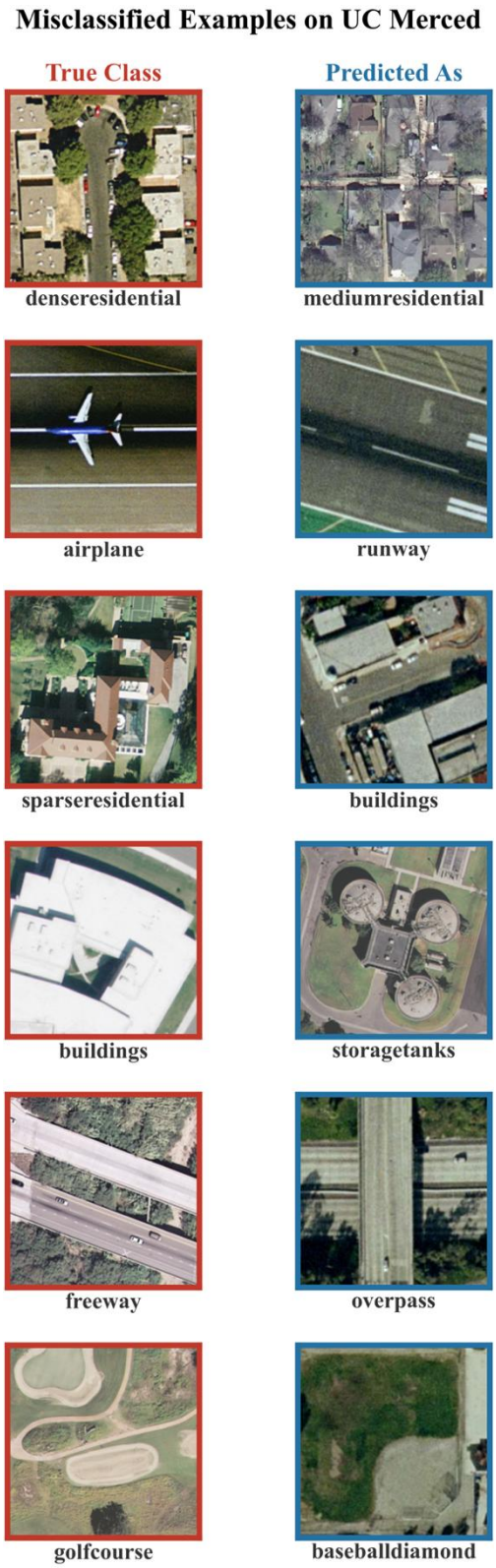


Figure 10. Misclassified examples from UC Merced

Future experiments using architecture-specific training configurations could clarify whether the observed accuracy ceiling changes when each model is optimized according to its own recommended training recipe. Our present data can only speak to the CNN side of this question: the five convolutional models converged to nearly the same accuracy under the shared protocol, which suggests that saturation on these benchmarks is not an artifact of the training configuration. For ViT-B/16, however, the degree to which a stronger training configuration would close the gap with Swin-T and ConvNeXt-Tiny remains an open empirical research question.

**C. Parameter Efficiency**

The efficiency results reported in Table 6 suggest that adding more parameters does not translate to higher accuracy on these benchmarks. Increasing parameters from 4.0M (EfficientNet-B0) to 85.8M (ViT-B/16) produced no accuracy gain on EuroSAT and a 1.42 percentage point decrease on UC Merced. DenseNet-121 and EfficientNet-B3 fall between these extremes in size and performed comparably to both. This pattern is consistent with the statistical analysis in Results, where most pairwise differences were not significant. Under memory or compute constraints, the model with the fewest parameters is a reasonable default given its competitive accuracy. However, parameter count and total training time provide only a partial view of computational efficiency. Since inference latency, FLOPs, GPU memory consumption, and throughput were not measured, the efficiency comparison should be interpreted as a limited parameter- and training-time-based analysis.

**D. Benchmark Saturation**

The observation that all eight models exceed 98% on EuroSAT, and that seven of eight exceed 99% on UC Merced, raises questions about the continued utility of these datasets for differentiating modern architectures. The accuracy spread on EuroSAT (0.41 percentage points) and the predominantly non-significant McNemar p-values suggest that performance differences at this level are largely attributable to test-set sampling variability rather than to genuine architectural advantages. Cheng et al. [4] noted the saturation of UC Merced as early as 2017, and our results provide additional quantitative support for that assessment. This interpretation is consistent with prior reports showing that both UC Merced and EuroSAT have already reached very high accuracy levels in recent transfer-learning studies. Although this study does not provide a full historical meta-analysis, the narrow accuracy range observed here supports the view that the remaining performance margin on these benchmarks is still limited. An important direction for future investigation is to determine whether architecture choice becomes more consequential on less saturated benchmarks such as NWPU-RESISC45 (45 classes, higher visual diversity), BigEarthNet (multi-label Sentinel-2 patches), or fMoW (temporal satellite classification with 62 categories).

**E. Limitations**

First, every result in this paper comes from one stratified 80/20 partition generated with seed 42. Because the experiments were conducted on a single data split, no standard deviations or confidence intervals are reported for OA, F1-Macro, and Cohen's kappa. Thus, the reported metrics should be interpreted as single-split estimates rather than variability-aware performance measures. McNemar's test can assess whether two models disagree on individual images within that partition, but it cannot tell us whether the overall ranking would survive a different data split. Running the experiment with several value random seeds and reporting means with standard deviations would provide a stronger basis for comparison; without such repetition, the rankings reported here are best regarded as indicative.

Second, the same test set was used both for early stopping and for final performance reporting, as no separate validation partition was held out. Consequently, the reported test accuracies should be interpreted as indicative estimates under the present evaluation protocol rather than definitive measurements of generalization performance. Current design may slightly overestimate model performance because the same partition influenced training termination and final reporting. Therefore, the comparative conclusions drawn in this study are limited to the selected datasets, split, and evaluation procedure. This is a standard practice in scene classification studies, but it creates an optimistic bias because the model selection criterion is evaluated on the data subsequently used to measure generalization performance.

Third, as discussed in Section V.B, the uniform training protocol is closer in configuration to standard CNN fine-tuning than to the heavy-augmentation regimes recommended for vision transformers. The poor performance of ViT-B/16 relative to other models may therefore reflect training configuration incompatibility rather than an inherent limitation of global self-attention

Fourth, only the RGB bands (B4, B3, B2) of EuroSAT were used, discarding 10 of the 13 available spectral channels. Although this choice ensured compatibility with ImageNet-pretrained models, it limits the generalizability of the findings to multispectral remote sensing workflows. Therefore, the conclusions should be interpreted within the RGB-only setting.

Fifth, all models were pretrained on ImageNet-1K with supervised learning. A larger pretraining dataset such as ImageNet-21K, or self-supervised methods built for satellite imagery, could shift which model performs best. For example, SatMAE learns representations from multi-temporal Sentinel-2 scenes using masked image modeling, while SeCo (Mañas et al., 2021) exploits seasonal variation in repeat-pass satellite imagery as a contrastive learning signal.

Finally, both EuroSAT and UC Merced are widely regarded as saturated benchmarks. The convergence of architectures observed here may not persist on more challenging datasets with larger class vocabularies or multi-label annotation. On the efficiency side, we measured only parameter counts and total training time. Per-epoch compute cost, FLOPs per forward pass, and actual inference throughput on target hardware were not recorded, which limits the practical conclusions that can be drawn about deployment readiness.

## CONCLUSIONS AND SUGGESTIONS

### Conclusion

This study evaluated eight deep learning architectures from three design families (classical CNNs, vision transformers, and a modernized CNN) on EuroSAT and UC Merced under a strictly uniform training protocol. With ImageNet-1K pretraining and identical optimization settings, every model surpassed 98% overall accuracy for both of benchmarks. ConvNeXt-Tiny obtained the highest accuracy on EuroSAT (99.11%) and Swin-T on UC Merced (99.76%); however, McNemar's test indicated that the majority of pairwise accuracy differences on EuroSAT were not statistically significant at alpha = 0.05. The total accuracy range was 0.41 percentage points on EuroSAT and 1.66 on UC Merced.

Under this specific protocol (one CNN-oriented training configuration, one data partition, ImageNet-1K supervised pretraining, and two saturated benchmarks), the evaluated architectures showed only marginal performance differences. Therefore, the findings should not be interpreted as evidence that architecture choice is generally less important than training configuration, but rather as evidence that these architectures converged under the selected experimental setting. EfficientNet-B0, the smallest experiment model evaluated (4.0M parameters), achieved accuracies within one percentage point of all other architectures on both datasets. This convergence is consistent with observations by Wightman et al. [19] on ImageNet, though we note that our experiment tests only one side of the comparison: we varied architecture while fixing the configuration, but did not vary the configuration while fixing the architecture, which would be needed to directly quantify the relative contribution of each factor.

### Suggestion

These conclusions rest on a single data partition, a CNN-oriented training configuration, RGB-only input, and two benchmarks that are widely regarded as saturated. Follow-up studies should evaluate the same architectures on NWPU-RESISC45 or BigEarthNet dataset and repeat the experiments with multiple random seeds to obtain variance estimates. Separately, comparing uniform and architecture-specific training configurations side by side would clarify how much the protocol itself affects rankings. Testing remote-sensing-specific pretrained weights from SatMAE or SeCo could also help reveal whether domain-specific pretraining shifts which model performs best.

## REFERENCES

Adegun, A. A., Viriri, S., & Tapamo, J. R. (2023). Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis. *Journal of Big Data*, *10*, 93. https://doi.org/10.1186/s40537-023-00772-x

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., & Khan, F. S. (2023). Transformers in Remote Sensing: A Survey. *Remote Sensing*, *15*(7), 1860. https://doi.org/10.3390/rs15071860

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

Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: benchmark and state of the art. *Proceedings of the IEEE*, *105*(10), 1865–1883.

https://doi.org/10.1109/JPROC.2017.2675998

Cheng, G., Xie, X., Han, J., Guo, L., & Xia, G.-S. (2020). Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 3735–3756. https://doi.org/10.1109/JSTARS.2020.3005403

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35–46. https://doi.org/10.1016/0034-4257(91)90048-B

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. https://arxiv.org/abs/2010.11929

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*(1), 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4

Foody, G. M. (2004). Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering & Remote Sensing*, *70*(5), 627–633. https://doi.org/10.14358/PERS.70.5.627

Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., Chellappa, R., Wilson, A. G., & Goldstein, T. (2023). Battle of the Backbones: A Large-Scale Comparison of Pretrained Models across Computer Vision Tasks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 29343–29371). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/5d9571470bb750f0e2325a030016f63f-Paper-Datasets_and_Benchmarks.pdf

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 (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). EuroSAT: a novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(7), 2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., & Chanussot, J. (2022). SpectralFormer: rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–15. https://doi.org/10.1109/TGRS.2021.3130716

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(6), e1264. https://doi.org/10.1002/widm.1264

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on*

*Computer Vision and Pattern Recognition*, 11976–11986.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & 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

Mañas, O., Lacoste, A., Giró-i-Nieto, X., Vazquez, D., & Rodriguez, P. (2021). Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9414–9423. https://doi.org/10.1109/ICCV48922.2021.00928

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

Neumann, M., Pinto, A. S., Zhai, X., & Houlsby, N. (2019). In-domain representation learning for remote sensing. *ArXiv Preprint ArXiv:1911.06721*.

Neyshabur, B., Sedghi, H., & Zhang, C. (2020). What is Being Transferred in Transfer Learning? *Advances in Neural Information Processing Systems (NeurIPS)*, *33*.

Nogueira, K., Penatti, O. A. B., & dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, *61*, 539–556. https://doi.org/10.1016/jß.patcog.2016.07.001

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

Wang, D., Zhang, Q., Xu, Y., Zhang, J., & Zhong, Y. (2023). Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–15. https://doi.org/10.1109/TGRS.2022.3222818

Wightman, R., Touvron, H., & Jégou, H. (2021). ResNet strikes back: an improved training procedure in timm. *ArXiv Preprint ArXiv:2110.00476*.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., & Lu, X. (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

Yang, Y., & Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 270–279. https://doi.org/10.1145/1869790.1869829

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (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