Comparative Analysis of Deep Learning Architectures for Land Cover Classification Using Sentinel-2 Imagery: A Case Study of Jambi Province, Indonesia

First Author, Second Author, and Third Author,

***Abstract*—This study evaluates five deep learning architectures for pixel-based land cover classification in Jambi Province, Sumatra, Indonesia, using multi-spectral Sentinel-2 imagery at 20-m resolution. The input feature space comprises 23 channels constructed from 10 spectral bands and 13 derived spectral indices. Ground truth labels were obtained from the Indonesian Ministry of Environment and Forestry (KLHK) 2024 land cover dataset covering 28,100 polygons, mapped to six simplified classes: Water, Trees/Forest, Crops/Agriculture, Shrub/Scrub, Built Area, and Bare Ground. Five architectures were compared: Swin Transformer Tiny (Swin-T), ConvNeXt Tiny (ConvNeXt-T), DenseNet-121, ResNet-50, and EfficientNet-B3, all initialised with ImageNet pretrained weights and fine-tuned on 80,000 training patches of 32×32 pixels. Swin-T achieved the highest overall accuracy of 79.54% with an F1-macro of 0.5644 and a Cohen's kappa of 0.587, followed by ConvNeXt-T at 79.14%, DenseNet-121 at 78.85%, ResNet-50 at 78.64%, and EfficientNet-B3 at 78.29%. McNemar's tests confirmed that Swin-T significantly outperformed all other architectures at $p < 0.05$. Per-class analysis revealed that all models achieved high F1 scores for Crops (0.839–0.846) but struggled with minority classes such as Bare Ground (0.207–0.278) and Shrub (0.174–0.367), reflecting severe class imbalance in tropical land cover distributions. The results demonstrate that modern transformer-based and convolution-based architectures offer statistically significant improvements over traditional residual networks for tropical land cover mapping.**

***Index Terms*—Land cover classification, deep learning, Sentinel-2, transfer learning, Swin Transformer, ConvNeXt, Jambi Province, Indonesia.**

I. Introduction

ACCURATE and up-to-date land cover maps are essential for environmental monitoring, natural resource management, and climate change assessment [1]. In tropical regions such as Sumatra, Indonesia, rapid land use change driven by palm oil expansion, logging, and urbanisation has made land cover mapping both urgent and challenging [2]. Jambi Province, located in central Sumatra, exemplifies these dynamics: its landscape spans lowland tropical forest, extensive oil palm and rubber plantations, peatland ecosystems, and growing urban centres, creating a heterogeneous classification problem that demands robust methods [3].

Manuscript received XXX; revised XXX.

First Author is with the Department, University, City, Country (e-mail: email@university.ac.id).

Second Author and Third Author are with the Department, University, City, Country.

Satellite remote sensing provides the spatial and temporal coverage needed for large-area land cover mapping. The European Space Agency's Sentinel-2 mission delivers multi-spectral imagery at 10 to 20 m resolution with a five-day revisit cycle, making it a primary data source for land cover studies [4]. Sentinel-2's spectral bands spanning visible, near-infrared, red edge, and shortwave infrared wavelengths enable discrimination among vegetation types, water bodies, built-up areas, and bare surfaces [5]. Spectral indices such as the Normalised Difference Vegetation Index (NDVI) and Normalised Difference Water Index (NDWI) further enhance class separability [6].

Traditional machine learning classifiers, including Random Forest and Support Vector Machines, have been widely applied to satellite image classification [7], [8]. However, these methods typically operate on individual pixels without exploiting spatial context. Deep learning architectures, particularly Convolutional Neural Networks (CNNs), have demonstrated superior performance by learning hierarchical spatial-spectral features directly from image patches [9], [10]. Transfer learning from ImageNet pretrained models further reduces data requirements [11].

The evolution of deep learning architectures has been rapid. Residual Networks (ResNet) introduced skip connections enabling training of very deep networks [12]. DenseNet improved feature reuse through dense connectivity [13]. EfficientNet achieved state-of-the-art performance through compound scaling [14]. More recently, the Swin Transformer incorporated hierarchical representations with shifted window attention [15], while ConvNeXt revisited pure convolutional designs using modern training strategies inspired by transformers [16].

Despite these advances, comparative studies of modern architectures for tropical land cover classification remain limited. Most studies focus on temperate environments [17], use coarser-resolution imagery [18], or compare only CNN variants without transformer-based models [19]. Land cover mapping in Indonesia faces specific challenges: persistent cloud cover, spectral similarity between plantation agriculture and forest, and severe class imbalance [20].

This study addresses these gaps by comparing five modern deep learning architectures for land cover classification in Jambi Province using cloud-free Sentinel-2 imagery. The architectures span three design paradigms: traditional CNNs

(ResNet-50, DenseNet-121), efficient CNNs (EfficientNet-B3, ConvNeXt-T), and vision transformers (Swin-T). The comparison covers overall accuracy, per-class metrics, training dynamics, and statistical significance through McNemar's pairwise tests.

II. Related Work

A. Deep Learning for Remote Sensing Classification

Deep learning has transformed remote sensing image classification. Early applications adapted CNN architectures to multi-spectral classification tasks [9]. Hu et al. [21] demonstrated that CNNs could effectively classify hyperspectral pixels, while subsequent work showed that 2D-CNNs capture spatial-spectral features jointly [10]. ResNet [12] addressed the vanishing gradient problem and enabled training beyond 100 layers, with gains transferring to remote sensing [22]. DenseNet [13] promoted feature reuse with fewer parameters. EfficientNet [14] established a new efficiency frontier through compound scaling, and several remote sensing studies have adopted it for land cover classification [23].

B. Vision Transformers for Image Classification

The Vision Transformer (ViT) [24] applied self-attention to image patches, demonstrating strong performance on large datasets. However, its quadratic complexity with image size poses challenges for remote sensing [25]. The Swin Transformer [15] addressed these limitations through hierarchical architecture with shifted window self-attention, achieving linear complexity. In remote sensing, Swin Transformer has shown promising results for land cover mapping [26]. ConvNeXt [16] modernised ResNet with transformer-inspired design choices, matching Swin Transformer performance while retaining convolutional inductive biases.

C. Transfer Learning for Remote Sensing

Transfer learning from ImageNet pretraining provides generic features that transfer to satellite imagery despite the domain gap [11], [27]. The primary adaptation challenge for Sentinel-2 is the channel mismatch: ImageNet models expect three-channel input, whereas Sentinel-2 provides 10 or more bands [28]. Common strategies include replicating pretrained weights across additional channels, or using the `timm` library [29] for automatic adaptation.

D. Land Cover Classification in Indonesia

Indonesia's tropical landscapes present particular classification challenges. The KLHK maintains an official land cover dataset serving as the national reference for forest monitoring [30]. Several studies have applied machine learning to land cover mapping in Sumatra [7], [31], but deep learning comparisons using modern architectures and Sentinel-2 data for Indonesian provinces remain limited.

TABLE I
KLHK Land Cover Code Mapping to Simplified Classes

ID	Class Name	KLHK Categories
0	Water	Tubuh Air (water bodies)
1	Trees/Forest	Hutan Lahan Kering Primer/Sekunder, Hutan Rawa, Hutan Mangrove, Hutan Tanaman
4	Crops/Agriculture	Pertanian Lahan Kering, Sawah, Perkebunan
5	Shrub/Scrub	Semak/Belukar, Semak Rawa
6	Built Area	Pemukiman (settlement)
7	Bare Ground	Tanah Terbuka, Pertambangan

III. Materials and Method

A. Study Area

Jambi Province is located on the eastern coast of Sumatra, Indonesia, between approximately $0.45°$ S to $2.45°$ S latitude and $101.1°$ E to $104.55°$ E longitude. The province covers approximately 50,160 km^2 and encompasses diverse land cover types ranging from lowland tropical rainforest in the western highlands to extensive oil palm and rubber plantations in the central lowlands and mangrove forests along the eastern coast. The tropical climate features a wet season from October to March and a drier period from April to September.

B. Data Sources

1) Sentinel-2 Satellite Imagery: Multi-spectral satellite imagery was acquired from the Copernicus Sentinel-2 mission via Google Earth Engine [32]. The `COPERNICUS/S2_SR_HARMONIZED` collection for 2024 was used, with cloud filtering via Cloud Score+ at a threshold of 0.60. A median composite was generated across the time period. The final product comprises four tiles totalling 2.7 GB. Ten spectral bands at 20 m resolution were retained: B2 (Blue, 490 nm), B3 (Green, 560 nm), B4 (Red, 665 nm), B5–B7 (Red Edge, 705–783 nm), B8 (NIR, 842 nm), B8A (Red Edge 4, 865 nm), B11 (SWIR 1, 1610 nm), and B12 (SWIR 2, 2190 nm).

2) KLHK Ground Truth Data: Ground truth labels were obtained from the Indonesian Ministry of Environment and Forestry (KLHK) 2024 land cover dataset (PL2024), downloaded in KMZ format via partitioned download across 29 spatial partitions. The dataset contains 28,100 polygons. Table I shows the mapping to six simplified classes.

C. Feature Engineering

Thirteen spectral indices were calculated from the 10 Sentinel-2 bands to enhance class separability. The indices span vegetation (NDVI, EVI, SAVI, MSAVI, GNDVI), water (NDWI, MNDWI), built-up (NDBI, BSI), red edge (NDRE, CIRE), and moisture (NDMI, NBR) domains. Table II presents the formulations.

The 10 spectral bands and 13 indices were stacked to produce a 23-channel feature raster. NaN and infinity values were replaced with zero, and each channel was independently normalised to zero mean and unit variance.

TABLE II
SPECTRAL INDICES CALCULATED FROM SENTINEL-2 BANDS

Index	Formula	Domain
NDVI	$\frac{\text{NIR}-\text{Red}}{\text{NIR}+\text{Red}}$	Vegetation
EVI	$2.5 \times \frac{\text{NIR}-\text{Red}}{\text{NIR}+6\text{R}-7.5\text{B}+1}$	Vegetation
SAVI	$1.5 \times \frac{\text{NIR}-\text{Red}}{\text{NIR}+\text{Red}+0.5}$	Vegetation
GNDVI	$\frac{\text{NIR}-\text{Green}}{\text{NIR}+\text{Green}}$	Vegetation
NDWI	$\frac{\text{Green}-\text{NIR}}{\text{Green}+\text{NIR}}$	Water
MNDWI	$\frac{\text{Green}-\text{SWIR1}}{\text{Green}+\text{SWIR1}}$	Water
NDBI	$\frac{\text{SWIR1}-\text{NIR}}{\text{SWIR1}+\text{NIR}}$	Built-up
BSI	$\frac{(\text{SWIR1}+\text{R})-(\text{NIR}+\text{B})}{(\text{SWIR1}+\text{R})+(\text{NIR}+\text{B})}$	Built-up
NDRE	$\frac{\text{NIR}-\text{RE1}}{\text{NIR}+\text{RE1}}$	Red Edge
CIRE	$\frac{\text{NIR}}{\text{RE1}} - 1$	Red Edge
NDMI	$\frac{\text{NIR}-\text{SWIR1}}{\text{NIR}+\text{SWIR1}}$	Moisture
NBR	$\frac{\text{NIR}-\text{SWIR2}}{\text{NIR}+\text{SWIR2}}$	Moisture

TABLE III
CLASS DISTRIBUTION IN THE TEST SET (20,000 PATCHES)

ID	Class Name	Samples	%
0	Water	223	1.12
1	Trees/Forest	7,416	37.08
2	Crops/Agriculture	11,470	57.35
3	Shrub/Scrub	35	0.18
4	Built Area	556	2.78
5	Bare Ground	300	1.50
	Total	**20,000**	**100.00**

D. Data Preparation

The KLHK polygons were rasterised onto the Sentinel-2 grid at 20 m resolution. A sliding window approach extracted 32×32 pixel patches with a stride of 16 pixels (50% overlap). Each patch was assigned the label of its centre pixel. A total of 100,000 patches were extracted via stratified sampling. The dataset was split into 80,000 training and 20,000 test samples using stratified random splitting (seed = 42). Table III shows the class distribution.

E. Model Architectures

Five deep learning architectures were selected representing three design paradigms. Table IV summarises their characteristics. All models were initialised with ImageNet pretrained weights. The first layer was modified to accept 23 input channels: for ResNet-50 and DenseNet-121, pretrained weights were averaged across the 3 RGB channels and replicated 23 times, scaled by $3/23$; for EfficientNet-B3, ConvNeXt-T, and Swin-T, the `timm` library [29] handled channel adaptation. The final classification layer was replaced with a linear layer mapping to 6 classes. All layers were trainable for end-to-end fine-tuning.

F. Training Protocol

All models were trained with identical hyperparameters: Adam optimiser with learning rate 10^{-4}, batch size 16, 30 epochs, cross-entropy loss with inverse-frequency class weights, ReduceLROnPlateau scheduler (factor 0.5, patience 3), and gradient clipping at max norm 1.0. Data augmentation during training comprised random horizontal flip ($p = 0.5$), random vertical flip ($p = 0.5$), and random rotation by $90°$, $180°$, or $270°$ ($p = 0.5$). Model selection was based on the highest validation accuracy, with the best checkpoint restored for test evaluation.

TABLE IV
SUMMARY OF DEEP LEARNING ARCHITECTURES

Architecture	Paradigm	Params (M)	Key Mechanism
ResNet-50	Residual CNN	23.6	Skip connections
DenseNet-121	Dense CNN	7.0	Dense connectivity
EfficientNet-B3	Efficient CNN	10.7	Compound scaling
ConvNeXt-T	Modern CNN	27.9	Large kernels, LayerNorm
Swin-T	Vis. Transformer	27.5	Shifted window attention

TABLE V
OVERALL CLASSIFICATION PERFORMANCE. BEST VALUES IN BOLD.

Model	OA (%)	F1-Ma	F1-Wt	Kappa	Params (M)	Time (min)
Swin-T	**79.54**	**0.5644**	**0.7836**	**0.587**	27.5	188.0
ConvNeXt-T	79.14	0.5469	0.7785	0.578	27.9	45.1
DenseNet-121	78.85	0.5390	0.7751	0.571	**7.0**	102.1
ResNet-50	78.64	0.5250	0.7727	0.566	23.6	96.1
EfficientNet-B3	78.29	0.5510	0.7685	0.557	10.7	218.3

G. Evaluation Metrics

Performance was assessed using overall accuracy (OA), F1-macro, F1-weighted, and Cohen's kappa coefficient [33], [34]. Per-class precision (user's accuracy), recall (producer's accuracy), and F1-score were computed. Statistical significance was assessed using McNemar's test [35] for all 10 pairwise comparisons at $\alpha = 0.05$.

IV. RESULTS AND DISCUSSION

A. Overall Classification Performance

Table V summarises the overall performance of all five architectures on the 20,000-sample test set. Swin-T achieved the highest accuracy of 79.54% with an F1-macro of 0.5644, F1-weighted of 0.7836, and kappa of 0.587. ConvNeXt-T ranked second at 79.14% accuracy, followed by DenseNet-121 at 78.85%, ResNet-50 at 78.64%, and EfficientNet-B3 at 78.29%. The accuracy spread across all five models is 1.25 percentage points.

The relatively low F1-macro values (0.5250–0.5644) compared with F1-weighted values (0.7685–0.7836) reflect poor performance on minority classes, which pull down the unweighted average. ConvNeXt-T achieved the fastest training time at 45.1 min, 4.2 times faster than Swin-T (188.0 min), while sacrificing only 0.40 percentage points in accuracy. DenseNet-121 offers the most parameter-efficient option at 7.0 M parameters yet trails the best accuracy by only 0.69 percentage points.

TABLE VI
PER-CLASS F1 SCORES. BEST VALUE PER CLASS IN BOLD.

Class	Swin-T	CNeXt	Dense	ResNet	EffNet
Water	0.624	0.626	0.596	**0.663**	0.583
Trees/Forest	**0.732**	0.724	0.719	0.713	0.708
Crops/Agri.	**0.846**	0.843	0.842	0.841	0.839
Shrub/Scrub	**0.367**	0.311	0.308	0.174	0.364
Built Area	0.554	**0.570**	0.543	0.546	0.435
Bare Ground	0.263	0.207	0.226	0.213	**0.278**

TABLE VII
PER-CLASS PRECISION (UA) AND RECALL (PA) FOR ALL ARCHITECTURES

Class		Swin-T	CNeXt	Dense	ResNet	EffNet
Water	UA	0.726	0.809	0.699	0.831	0.882
	PA	0.547	0.511	0.520	0.552	0.435
Trees	UA	0.904	0.907	0.894	0.899	0.914
	PA	0.616	0.603	0.602	0.591	0.577
Crops	UA	0.762	0.757	0.755	0.752	0.742
	PA	0.951	0.952	0.951	0.954	0.965
Shrub	UA	0.643	0.700	0.471	0.364	0.387
	PA	0.257	0.200	0.229	0.114	0.343
Built	UA	0.681	0.644	0.761	0.770	0.889
	PA	0.468	0.511	0.423	0.423	0.288
Bare	UA	0.671	0.638	0.583	0.417	0.573
	PA	0.163	0.123	0.140	0.143	0.183

B. Per-Class Analysis

Per-class F1 scores are presented in Table VI. Crops/Agriculture achieved the highest F1 scores across all models (0.839–0.846), followed by Trees/Forest (0.708–0.732). These two classes together constitute 94.43% of the test set. Bare Ground achieved the lowest F1 scores (0.207–0.278), followed by Shrub/Scrub (0.174–0.367).

Table VII presents precision (user's accuracy, UA) and recall (producer's accuracy, PA) per class. All models exhibit a strong precision–recall trade-off. Crops achieves consistently high recall (0.951–0.965) but moderate precision (0.742–0.762), indicating many non-crop pixels are misclassified as cropland. Trees shows high precision (0.894–0.914) but moderate recall (0.577–0.616), meaning pixels classified as forest are likely correct, but many forest pixels are missed.

C. Confusion Matrix Analysis

The normalised confusion matrices for all five architectures are shown in Fig. 1. The most common misclassification across all models is Trees being classified as Crops, accounting for approximately 2,800–3,100 misclassified samples per model out of 7,416 total Trees samples. This confusion reflects the spectral similarity between forest canopy and plantation crops such as oil palm and rubber at 20 m resolution. Bare Ground is predominantly misclassified as Crops (193–206 out of 300 samples) and Trees (36–57 samples).

D. Training Dynamics

Fig. 2 presents the training and validation loss and accuracy curves for all architectures over 30 epochs. All models exhibit rapid initial improvement in the first 5–10 epochs, followed by gradual convergence. Swin-T reached its best validation accuracy at epoch 27, indicating that the transformer architecture benefits from extended training. DenseNet-121 peaked at epoch 25. ResNet-50 achieved its best performance at epoch 18, suggesting the simpler residual architecture saturates earlier. EfficientNet-B3 peaked at epoch 19. The gap between training and validation accuracy remains moderate (approximately 1–3 percentage points at convergence), indicating adequate regularisation.

TABLE VIII
MCNEMAR'S TEST RESULTS FOR PAIRWISE COMPARISONS ($\alpha = 0.05$)

Comparison	χ^2	p-value	Sig.
Swin-T vs EfficientNet-B3	44.068	<0.001	***
Swin-T vs ResNet-50	29.075	<0.001	***
Swin-T vs DenseNet-121	17.019	<0.001	***
Swin-T vs ConvNeXt-T	6.537	0.011	*
ConvNeXt-T vs EfficientNet-B3	19.292	<0.001	***
ConvNeXt-T vs ResNet-50	8.629	0.003	**
ConvNeXt-T vs DenseNet-121	2.906	0.088	ns
DenseNet-121 vs EfficientNet-B3	8.915	0.003	**
DenseNet-121 vs ResNet-50	1.628	0.202	ns
EfficientNet-B3 vs ResNet-50	3.287	0.070	ns

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant

E. Statistical Significance

McNemar's test was applied to all 10 pairwise comparisons. Table VIII presents the results, and Fig. 3 shows the p-value matrix.

The results establish a partial ordering. Swin-T significantly outperforms all other architectures at $p < 0.05$, with the strongest separation from EfficientNet-B3 ($\chi^2 = 44.068$, $p < 0.001$) and the weakest from ConvNeXt-T ($\chi^2 = 6.537$, $p = 0.011$). ConvNeXt-T significantly outperforms EfficientNet-B3 ($p < 0.001$) and ResNet-50 ($p = 0.003$) but not DenseNet-121 ($p = 0.088$). The differences among DenseNet-121, ResNet-50, and EfficientNet-B3 are not statistically significant, placing them in a second performance tier.

F. Spatial Prediction Analysis

Fig. 4 presents the Sentinel-2 true-colour composite and the KLHK ground truth land cover map for Jambi Province. The dense forest cover in the western highlands, extensive plantation agriculture in the central and eastern lowlands, and urban areas around the provincial capital are clearly visible.

Fig. 5 shows the ResNet-50 prediction map at the province scale. The model correctly captures the broad spatial distribution of forest and cropland but produces a smoother map than the reference data, particularly along class boundaries. This smoothing effect is expected because the 32×32 pixel patch-based classification integrates spatial context over a 640×640 m window.

G. Discussion

The results demonstrate that modern architectures offer statistically significant but modest improvements over traditional residual networks. Swin-T's advantage of 0.90 percentage

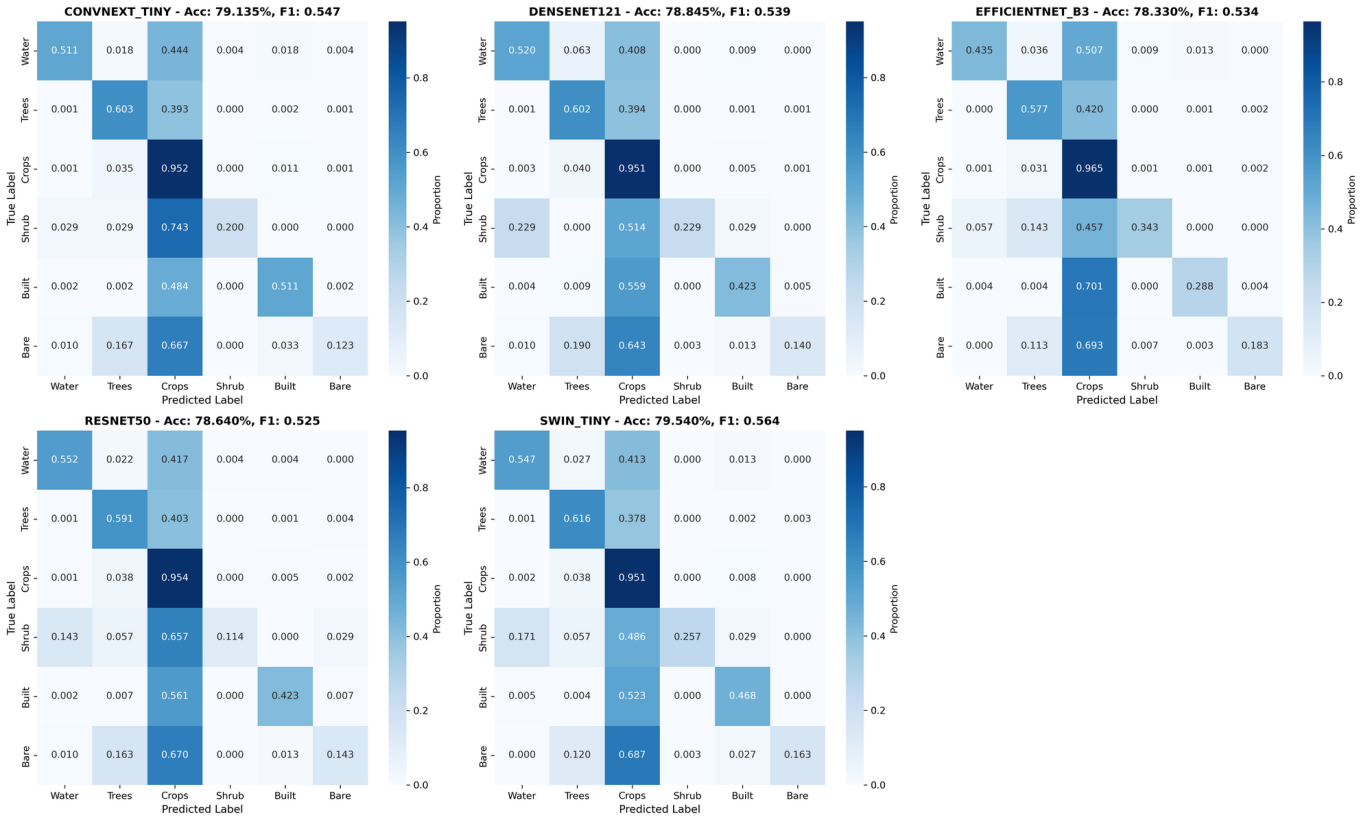

Fig. 1. Normalised confusion matrices for all five architectures on the 20,000-sample test set. Each cell shows the proportion of actual class samples (rows) classified as the predicted class (columns). Darker shading indicates higher proportions.

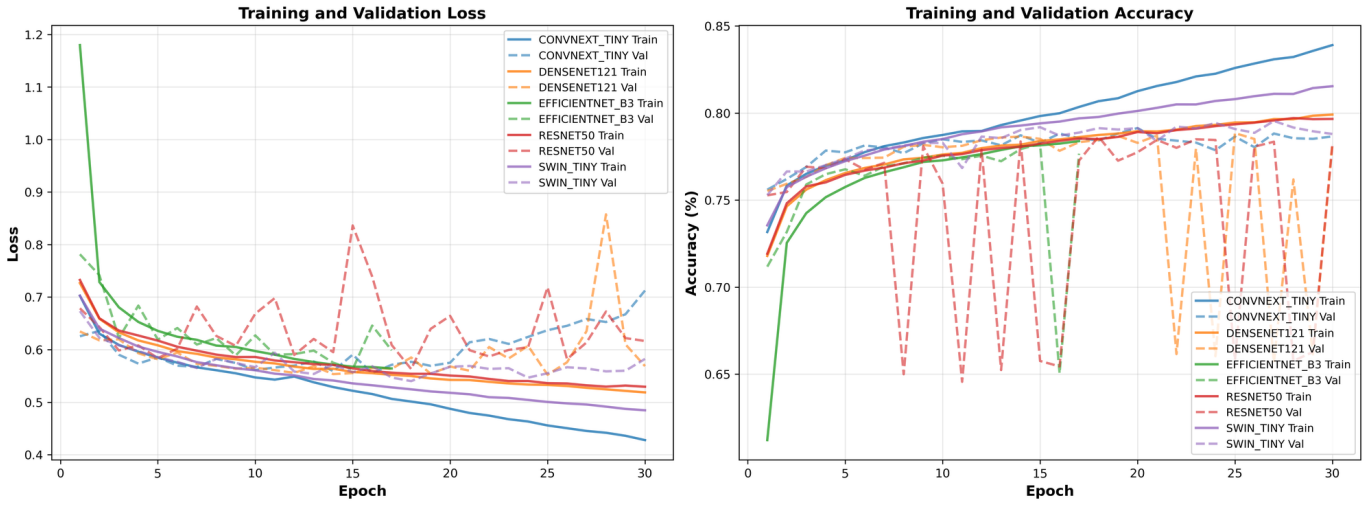

Fig. 2. Training and validation loss (left) and accuracy (right) curves for all five architectures over 30 training epochs. Solid lines represent training metrics; dashed lines represent validation metrics.

points over ResNet-50, while statistically significant ($p < 0.001$), translates to approximately 180 additional correctly classified patches out of 20,000.

The persistent challenge across all architectures is minority class classification. Bare Ground and Shrub, representing less than 2% of the dataset, achieve F1 scores below 0.40 regardless of architecture. This limitation stems from insufficient training samples and genuine spectral overlap with more common classes.

The comparison between Swin-T and ConvNeXt-T is informative. Despite representing different computational paradigms (self-attention versus convolution), these archi-

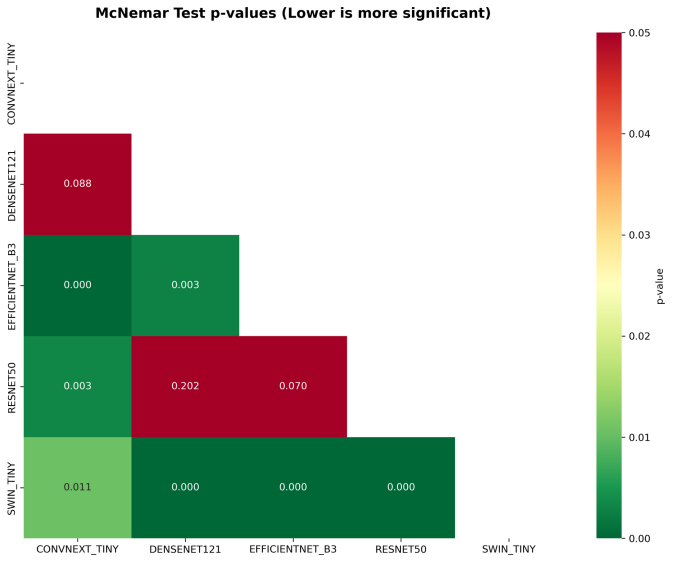

Fig. 3. McNemar's test p-value matrix. Green cells indicate statistically significant differences ($p < 0.05$); red cells indicate non-significant differences.

tectures achieve nearly identical accuracy (79.54% versus 79.14%) with similar parameter counts. ConvNeXt-T's 4.2× faster training makes it practical when computational resources are constrained, while Swin-T's marginally better F1-macro (0.5644 versus 0.5469) suggests self-attention may capture subtle patterns benefiting minority class recognition.

DenseNet-121 presents an attractive efficiency trade-off. With only 7.0 M parameters (one-quarter of Swin-T and ConvNeXt-T), it achieves 78.85% accuracy. For deployment scenarios where model size is critical, DenseNet-121 offers the best accuracy-per-parameter ratio.

The moderate kappa values (0.557–0.587) indicate that while the models outperform random classification, there is room for improvement. This is consistent with the spectral similarity between natural forest and plantation crops at 20 m resolution.

Several limitations should be noted. First, a single year of imagery was used, missing temporal dynamics. Second, the 20 m resolution limits detection of small-scale features. Third, the patch-based approach introduces boundary effects. Fourth, the KLHK reference data itself contains uncertainties from visual interpretation. Fifth, only five architectures were compared.

V. Conclusion

This study compared five deep learning architectures for land cover classification in Jambi Province, Indonesia, using 23-channel Sentinel-2 imagery and KLHK ground truth data across six land cover classes. Swin Transformer Tiny achieved the best overall accuracy of 79.54% with an F1-macro of 0.5644, statistically outperforming all other architectures at $p < 0.05$ in McNemar's pairwise tests. ConvNeXt Tiny ranked second at 79.14% with 4.2× faster training, making it preferred when computational efficiency is prioritised. DenseNet-121 offered the best parameter efficiency at 7.0 M parameters while achieving 78.85% accuracy.

All architectures struggled with minority classes, particularly Bare Ground (F1 = 0.207–0.278) and Shrub/Scrub (F1 = 0.174–0.367). The dominant confusion between Trees/Forest and Crops/Agriculture highlights the fundamental challenge of distinguishing natural forest from plantation crops at 20 m resolution.

Future work should address minority class performance through focal loss or class-balanced sampling, incorporate temporal features from multi-date composites, explore multi-scale architectures combining 10 m and 20 m bands, and integrate SAR data from Sentinel-1 for structural discrimination.

References

[1] P. Gong, J. Wang, L. Yu, Y. Zhao, Y. Zhao, L. Liang, Z. Niu, X. Huang, H. Fu, S. Liu *et al.*, "Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data," *International Journal of Remote Sensing*, vol. 34, no. 7, pp. 2607–2654, 2013.

[2] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland *et al.*, "High-resolution global maps of 21st-century forest cover change," *Science*, vol. 342, no. 6160, pp. 850–853, 2013.

[3] Wahyunto, S. Ritung, and H. Subagjo, "Peatland distribution in Sumatra and Kalimantan – explanation of its data sets including source of information, data constraints, data analysis, and use of the data," Wetlands International – Indonesia Programme, Bogor, Tech. Rep., 2004.

[4] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort *et al.*, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012.

[5] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: a review," *Remote Sensing*, vol. 12, no. 14, p. 2291, 2020.

[6] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the Great Plains with ERTS," in *Proceedings of the 3rd Earth Resources Technology Satellite-1 Symposium*, vol. 1, 1974, pp. 309–317.

[7] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: a review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[8] G. M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1335–1343, 2004.

[9] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: a meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, 2019.

[10] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 6, p. e1264, 2018.

[11] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "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*, vol. 13, pp. 4697–4713, 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[13] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[14] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.

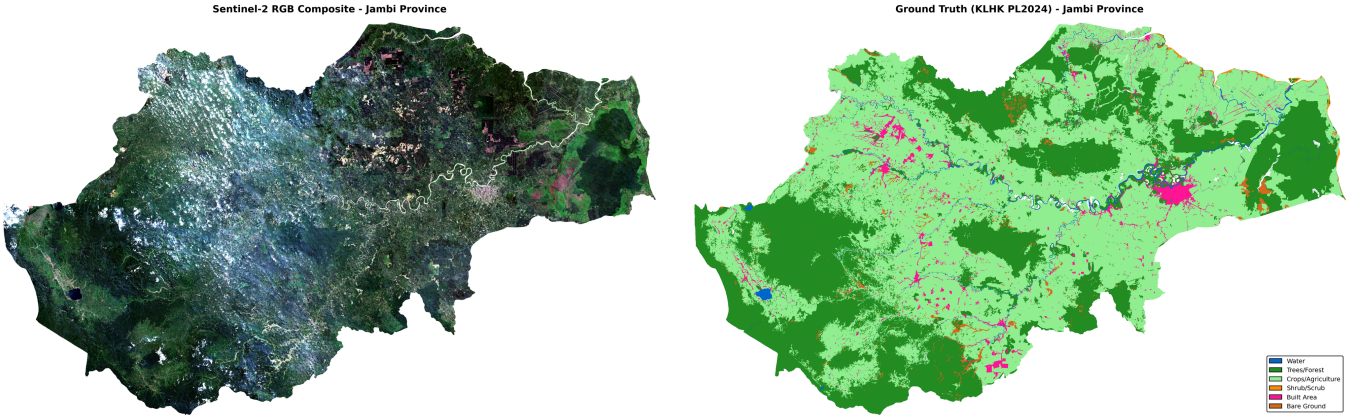

Fig. 4. Study area overview. Left: Sentinel-2 true-colour (RGB) composite of Jambi Province, 2024 dry season. Right: KLHK 2024 ground truth land cover map showing the six classified land cover types.

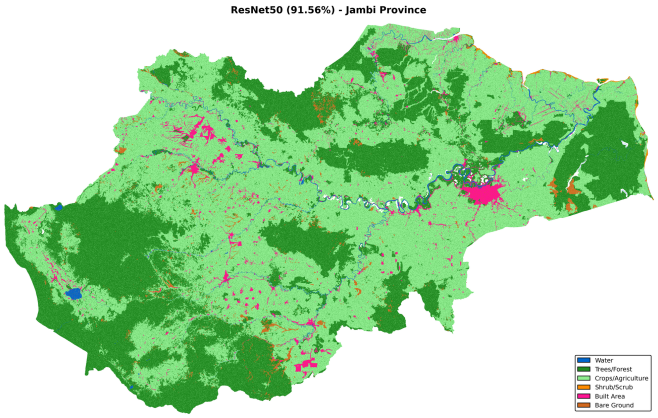

Fig. 5. ResNet-50 land cover prediction map for Jambi Province, demonstrating the spatial distribution of classified land cover types.

[16] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 966–11 976.

[17] M. Wessel, M. Brandmeier, and D. Tiede, "Evaluation of different machine learning algorithms for scalable classification of tree types and tree species based on Sentinel-2 data," *Remote Sensing*, vol. 10, no. 9, p. 1419, 2018.

[18] C. E. Woodcock, T. R. Loveland, M. Herold, and M. E. Bauer, "Transitioning from change detection to monitoring with remote sensing: a paradigm shift," *Remote Sensing of Environment*, vol. 238, p. 111558, 2020.

[19] D. Tong, Y. Zhang, and L. Zheng, "Land cover classification with multi-source data using evidential reasoning approach," *Chinese Geographical Science*, vol. 29, no. 5, pp. 799–812, 2019.

[20] E. L. Bullock, C. E. Woodcock, and P. Olofsson, "Monitoring tropical forest degradation using spectral unmixing and Landsat time series analysis," *Remote Sensing of Environment*, vol. 238, p. 110968, 2020.

[21] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, pp. 1–12, 2015.

[22] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: a comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[23] A. Naushad, T. Rahim, and M. Alam, "Analysis of Sentinel-2 MSI data using EfficientNet for land use/land cover classification," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 10, pp. 8355–8364, 2022.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[25] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[26] D. Wang, Q. Zhang, Y. Xu, J. Zhang, and Y. Zhong, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[27] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, "In-domain representation learning for remote sensing," *arXiv preprint arXiv:1911.06721*, 2019.

[28] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sensing of Environment*, vol. 80, no. 1, pp. 185–201, 2002.

[29] R. Wightman, "PyTorch Image Models," 2019, gitHub repository, https://github.com/huggingface/pytorch-image-models.

[30] Directorate General of Forestry Planning, "Indonesia national forest reference emission level for deforestation and forest degradation," Ministry of Environment and Forestry, Jakarta, Tech. Rep., 2016.

[31] J. Miettinen, C. Shi, and S. C. Liew, "Land cover distribution in the peatlands of Peninsular Malaysia, Sumatra and Borneo in 2015 with changes since 1990," *Global Ecology and Conservation*, vol. 6, pp. 67–78, 2016.

[32] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[34] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of Environment*, vol. 37, no. 1, pp. 35–46, 1991.

[35] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.