Cross-Dataset Generalization of ConvNeXt-Tiny for Remote Sensing Scene Classification

Nur Nafiiyah
Department of Informatics Engineering, Universitas Islam Lamongan, Lamongan, Indonesia
mynaff26@gmail.com (corresponding author)

Agus Harjoko
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia
aharjoko@ugm.ac.id

Achmad Nizar Hidayanto
Information Systems Department, Faculty of Computer Science, Universitas Indonesia, Jakarta, Indonesia
nizar@cs.ui.ac.id

Received: 17 January 2026 | Revised: 9 February 2026, 19 February 2026, 3 March 2026, and 4 March 2026 | Accepted: 6 March 2026

 DOI: https://doi.org/10.48084/etasr.17607

ABSTRACT

Remote Sensing Scene Classification (RSSC) is a fundamental task for understanding high-resolution aerial imagery and supports a wide range of applications such as land-use analysis, environmental monitoring, and urban planning. Despite recent advances in deep learning, many existing studies focus primarily on in-dataset evaluation, whereas the generalization capability of modern convolutional architectures under cross-dataset conditions remains insufficiently explored. To address this gap, this study investigates the effectiveness of ConvNeXt-Tiny as a transfer learning backbone for RSSC and systematically compares its performance with widely used Convolutional Neural Networks (CNNs), namely ResNet50, DenseNet121, and MobileNetV2. Experiments were conducted using two benchmark datasets, NWPU-RESISC45 and AID, with 20 shared scene categories. Four experimental scenarios were designed, including in-dataset evaluation on each dataset and cross-dataset evaluation without fine-tuning to assess robustness under domain shift. All models were pretrained on ImageNet and trained using an identical transfer learning protocol to ensure a fair comparison. Performance was evaluated using accuracy, precision, recall, F1-score, and macro-averaged Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Experimental results demonstrate that ConvNeXt-Tiny achieves strong in-dataset performance, matching or slightly outperforming ResNet50 on NWPU and showing competitive results on AID. More importantly, ConvNeXt-Tiny maintains robust cross-dataset generalization, achieving performance comparable to ResNet50 and significantly outperforming DenseNet121 and MobileNetV2. ROC-AUC analysis further confirms the stable discriminative capability of ConvNeXt-Tiny across different evaluation scenarios. These findings indicate that modern convolutional designs such as ConvNeXt-Tiny offer an effective and robust solution for RSSC, particularly under domain shift conditions.

Keywords-classification; ConvNeXt-Tiny; scene; remote sensing; transfer learning

I. INTRODUCTION

Remote Sensing Scene Classification (RSSC) plays a crucial role in various applications such as land-use and land-cover mapping [1], environmental monitoring, urban planning, and disaster risk assessment [2]. The rapid increase in high-resolution satellite [3] and aerial imagery has significantly intensified the demand for accurate and reliable scene classification methods [4]. However, RSSC remains challenging due to complex scene structures [5], scale variations, high inter-class similarity, and intra-class diversity [6].

In recent years, Convolutional Neural Networks (CNNs) have become the dominant approach for RSSC owing to their strong hierarchical feature representation capability [7]. Transfer learning with CNNs pretrained [8] on large-scale datasets such as ImageNet has demonstrated remarkable success in improving classification performance on benchmark datasets [9], including UC Merced [10], AID, and NWPU-RESISC45 [11]. To further enhance accuracy, many studies

have introduced architectural extensions [12] such as attention mechanisms [13], multi-branch structures [14], graph-based learning [15], and hybrid CNN-Transformer models [16].

Despite their effectiveness, existing approaches exhibit two notable limitations. First, most studies primarily emphasize in-dataset performance, whereas the generalization ability across different remote sensing datasets remains underexplored [17]. Given the substantial differences in spatial resolution, scene composition [18], and acquisition conditions among datasets, models trained on a single dataset often fail to generalize well to unseen domains [19]. Second, recent research trends increasingly focus on architectural complexity [20], which leads to higher computational cost, reduced reproducibility, and limited practical applicability [21].

ConvNeXt has recently emerged as a modernized CNN architecture that integrates design principles inspired by transformers while preserving the efficiency and simplicity of convolutional networks [22]. Although ConvNeXt has shown promising performance in various computer vision tasks, existing RSSC studies mainly employ ConvNeXt-Tiny as a backbone augmented with additional attention modules or hybrid components. Consequently, the intrinsic capability of ConvNeXt as a pure CNN transfer learning backbone, particularly its cross-dataset generalization behavior, has not been systematically investigated.

To address these gaps, this paper presents a comprehensive empirical evaluation of ConvNeXt-Tiny for RSSC under consistent and reproducible experimental settings. Unlike prior works that emphasize architectural augmentation, we focus on analyzing ConvNeXt as a standalone convolutional backbone within a standard supervised learning framework. Experiments are conducted on two widely used benchmark datasets, NWPU-RESISC45 and AID, under both in-dataset and cross-dataset evaluation scenarios. Furthermore, ConvNeXt-Tiny is compared fairly with other modern CNN architectures to assess its relative effectiveness and robustness.

The main contributions of this work are summarized as follows:

1. We provide a systematic evaluation of ConvNeXt-Tiny as a pure CNN transfer learning backbone for RSSC without incorporating additional attention or hybrid modules.
2. We investigate the cross-dataset generalization capability of ConvNeXt-Tiny using different training and testing dataset combinations.
3. We perform a fair comparative analysis against other state-of-the-art CNN architectures using consistent datasets, training strategies, and evaluation metrics, including accuracy, Area Under the Receiver Operating Characteristic Curve (ROC-AUC), and confusion matrix analysis.

II. RELATED WORK

RSSC has evolved significantly from handcrafted feature-based methods to deep learning-based approaches [23]. CNN architectures such as VGG, ResNet, DenseNet, and MobileNet have been widely adopted through transfer learning to extract discriminative scene features [24]. To further enhance performance [25], attention mechanisms, multi-scale feature fusion [26], and graph-based learning frameworks have been integrated into CNN backbones [27].

More recently, transformer-based and hybrid CNN-Transformer models have been introduced to capture long-range contextual information [28] and improve scene representation [29]. While these approaches achieve competitive results, they often involve complex architectures, high computational cost, and limited reproducibility [30].

ConvNeXt represents a modern CNN architecture inspired by transformer design principles. Existing RSSC studies primarily employ ConvNeXt-Tiny with additional architectural enhancements to improve in-dataset accuracy. However, a systematic evaluation of ConvNeXt as a standalone convolutional backbone, particularly in cross-dataset scenarios, remains limited. This study addresses this research gap by focusing on the generalization behavior of ConvNeXt-Tiny under consistent transfer learning settings [22].

III. DATASET

In this study, two widely used benchmark datasets for RSSC were employed, namely NWPU-RESISC45 and AID. These datasets are widely used for their size, diversity, and availability. The NWPU-RESISC45 dataset is available at [31] and is discussed in the study [23]. The AID dataset is available at [32] and is discussed in the study [26].

The NWPU-RESISC45 dataset consists of 31,500 remote sensing images categorized into 45 scene classes, with 700 images per class. These images were collected from diverse geographical regions and exhibit significant variations in scale, illumination, and spatial resolution, which makes the dataset highly challenging for scene classification tasks.

The AID dataset contains 10,000 aerial images distributed across 30 scene categories. The images were acquired from different countries and sensors, resulting in substantial diversity in scene composition and spatial resolution. Due to its heterogeneous characteristics, the AID dataset is commonly used to evaluate the robustness and generalization capability of RSSC models.

These two datasets exhibit distinct characteristics in terms of the number of classes, spatial resolution, and scene complexity, making them suitable for both in-dataset and cross-dataset evaluations.

To enable a fair cross-dataset comparison, a common subset of 20 shared scene classes was selected from both datasets. The selected classes include: airport, beach, bridge, church, commercial area, dense residential, desert, farmland, forest, industrial area, meadow, medium residential, mountain, park, parking lot, railway station, river, sparse residential, stadium, and storage tank. As a result, the final dataset comprised 14,000 images from NWPU and 6,700 images from AID. The training and testing splits for the NWPU and AID datasets are summarized in Table I.

TABLE I. DATASET DESCRIPTION

No	Class	AID			NWPU		
		Train	Test	Total	Train	Test	Total
1	airport	300	60	360	560	140	700
2	beach	300	100	400	560	140	700
3	bridge	300	60	360	560	140	700
4	church	200	40	240	560	140	700
5	commercial	300	50	350	560	140	700
6	dense_residential	300	110	410	560	140	700
7	desert	250	50	300	560	140	700
8	farmland	300	70	370	560	140	700
9	forest	200	50	250	560	140	700
10	industrial	300	90	390	560	140	700
11	meadow	240	40	280	560	140	700
12	medium_residential	250	40	290	560	140	700
13	mountain	300	40	340	560	140	700
14	park	300	50	350	560	140	700
15	parking	300	90	390	560	140	700
16	railway_station	220	40	260	560	140	700
17	river	300	110	410	560	140	700
18	sparse_residential	250	50	300	560	140	700
19	stadium	250	40	290	560	140	700
20	storage_tank	300	60	360	560	140	700
	Total	5,460	1,240	6,700	11,200	2,800	14,000

IV. MODEL CLASSIFICATION

ConvNeXt-Tiny was selected as the primary model in this study. ConvNeXt is a modern CNN that integrates design principles inspired by vision transformers while preserving the efficiency and structural simplicity of conventional CNNs. In this work, a ConvNeXt-Tiny model pretrained on ImageNet was adopted as the backbone feature extractor. The original classification head was removed and replaced with a global average pooling layer followed by fully connected layers and a softmax classifier corresponding to the number of scene classes. To focus on evaluating ConvNeXt-Tiny as a pure convolutional transfer learning backbone, no additional attention modules, transformer blocks, or hybrid components were incorporated into the network. For a fair and comprehensive comparison, three widely used CNN architectures were selected as baseline models:

- ResNet50, representing deep residual learning,
- DenseNet121, emphasizing feature reuse through dense connectivity,
- MobileNetV2, designed for computational efficiency and lightweight deployment.

All baseline models were pretrained on ImageNet and adapted to the RSSC task by replacing their final classification layers with task-specific fully connected layers. For all models, transfer learning was performed by freezing the convolutional backbone during the initial training stage and training only the newly added classification layers. This strategy reduces the risk of overfitting and ensures a fair comparison across different architectures. The same training protocol and hyperparameter settings were applied consistently to all models. The overall research workflow is illustrated in Figure 1.

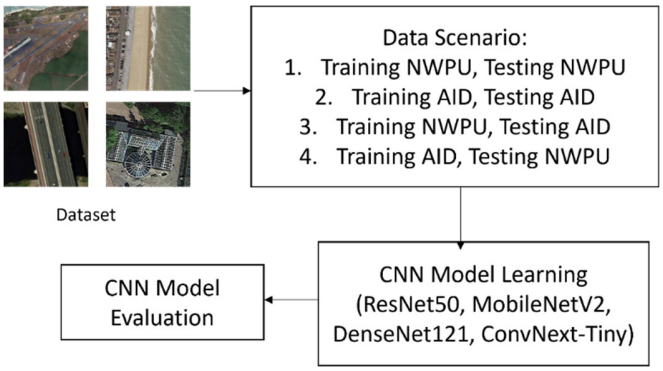


Fig. 1. Research workflow.

To comprehensively evaluate both classification performance and generalization capability, two experimental settings were designed: in-dataset evaluation and cross-dataset evaluation. In the in-dataset setting, models were trained and tested on the same dataset using a standard training-validation-testing split:

- NWPU-RESISC45 was used for both training and testing,
- AID was used for both training and testing.

This setting serves as a baseline to verify the effectiveness of each model under controlled and consistent data conditions.

In the cross-dataset setting, models were trained on one dataset and directly evaluated on the other dataset without any additional fine-tuning. Two cross-dataset scenarios were considered:

- Training on NWPU-RESISC45 and testing on AID,
- Training on AID and testing on NWPU-RESISC45.

Due to GPU memory limitations when training high-resolution remote sensing images with pretrained deep CNN backbones, the batch size was set to 2. Larger batch sizes resulted in memory overflow under the available hardware configuration. Despite the small batch size, the number of update steps per epoch remained sufficiently large (e.g., approximately 5,600 updates per epoch for NWPU-RESISC45 and 2,730 for AID), ensuring adequate gradient updates during training. To maintain stable optimization, the Adam optimizer with a learning rate of 1×10^{-4} and early stopping were employed. The same batch size was consistently applied across all models to guarantee a fair comparison.

This setting aims to assess the robustness and generalization capability of the models under domain shift conditions, where differences in image resolution, scene composition, and data distribution exist. All experiments were implemented using the TensorFlow framework. The categorical cross-entropy loss function was used for multi-class classification. All input images were resized to 224 × 224 pixels to match the ImageNet pretraining configuration. Models were trained for 30 epochs with a batch size of 2. To ensure fairness and reproducibility, identical training configurations and hyperparameters were consistently applied to all compared models.

Model performance was evaluated using multiple metrics to provide a comprehensive assessment: accuracy (1), precision (2), recall (3), and F1-score (4) [7].

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1} - \text{score} = 2 * \frac{\text{Precision*Recall}}{\text{Precision+Recall}} \quad (4)$$

In addition, the ROC-AUC was computed using a macro-averaging strategy to assess the overall discriminative ability of the models across all classes. Confusion matrices were also analyzed to examine class-wise prediction behavior and misclassification patterns. To ensure reproducibility, all experiments were conducted under consistent hardware and software environments. The same preprocessing procedures, data splits, and training protocols were applied to all models. Random seeds were fixed during training to reduce stochastic variability.

Due to the limited size of the datasets, formal statistical significance tests were not conducted; however, consistent performance trends across multiple metrics and experimental scenarios provide reliable comparative insights.

V. RESULTS AND DISCUSSION

Table II presents the evaluation results under Scenario 1, where all models were trained and tested on the NWPU-RESISC45 dataset using a standard train-test split. The results show that ConvNeXt-Tiny achieved the highest overall performance, obtaining an F1-score of 90.92% and an accuracy of 90.89%. It also achieved strong precision (91.39%) and recall (90.64%), indicating a well-balanced classification capability across the selected scene categories. ResNet50 achieved comparable performance, with an F1-score of 90.80% and an accuracy of 90.82%. The slight performance gap suggests that while deep residual learning is effective for complex scene representation, ConvNeXt-Tiny is able to extract equally discriminative features with a more modern convolutional design.

TABLE II. EVALUATION RESULTS OF SCENARIO 1

Model	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
ResNet50	90.80	90.82	91.74	90.43
MobileNetV2	63.81	64.68	72.38	59.14
DenseNet121	74.32	74.54	79.24	69.93
ConvNeXt-Tiny	90.92	90.89	91.39	90.64

In contrast, DenseNet121 achieved moderate performance, with an F1-score of 74.32%, indicating that dense feature reuse alone may not be sufficient to capture the high intra-class variability and inter-class similarity present in the NWPU dataset. MobileNetV2 obtained the lowest performance, with an F1-score of 63.81%, highlighting the limitations of lightweight architectures when applied to complex RSSC tasks. Overall, the results in Table II demonstrate that ConvNeXt-Tiny is highly effective for in-dataset classification on NWPU-RESISC45, outperforming or matching deeper CNN architectures while significantly exceeding the performance of lightweight models.

Table III presents the evaluation results when all models were trained and tested on the AID dataset. As shown in the table, ResNet50 achieved the best overall performance, obtaining an accuracy of 90.16% and an F1-score of 89.90%. ConvNeXt-Tiny demonstrated a very competitive performance, with an accuracy of 89.52% and an F1-score of 89.00%, showing only a marginal difference compared to ResNet50. In contrast, MobileNetV2 and DenseNet121 performed substantially worse on this scenario. MobileNetV2 achieved an accuracy of 64.52%, whereas DenseNet121 reached 68.79%, indicating that lightweight architectures and densely connected networks may struggle to capture the complex spatial patterns present in the AID dataset.

TABLE III. EVALUATION RESULTS OF SCENARIO 2

Model	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
ResNet50	89.90	90.16	90.91	89.52
MobileNetV2	63.39	64.52	71.40	60.81
DenseNet121	67.33	68.79	73.63	65.08
ConvNeXt-Tiny	89.00	89.52	90.65	89.11

Table IV reports the cross-dataset evaluation results when models were trained on NWPU-RESISC45 and tested on AID. As expected, all models experienced performance degradation compared to the in-dataset setting due to domain shift between the two datasets. ResNet50 achieved the highest accuracy of 80.75%, followed closely by ConvNeXt-Tiny with an accuracy of 80.31% and an F1-score of 78.63%. The small performance gap between the two models suggests that ConvNeXt-Tiny maintains strong generalization capability, even when evaluated on unseen data distributions. In contrast, MobileNetV2 and DenseNet121 showed a severe drop in performance, with F1-scores of 43.07% and 55.35%, respectively. This indicates that these architectures are less robust to cross-dataset variations in scene composition, resolution, and acquisition conditions.

TABLE IV. EVALUATION RESULTS OF SCENARIO 3

Model	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
ResNet50	79.23	80.75	83.02	79.82
MobileNetV2	43.07	47.07	53.37	44.04
DenseNet121	55.35	57.33	61.95	55.03
ConvNeXt-Tiny	78.63	80.31	81.78	79.85

The results of Scenario 4, shown in Table V, further confirm the challenges of cross-dataset generalization. When trained on AID and tested on NWPU-RESISC45, all models exhibited lower performance compared to Scenario 3, reflecting the higher scene diversity and complexity of NWPU. ResNet50 achieved the highest accuracy of 73.87%, whereas ConvNeXt-Tiny closely followed with an accuracy of 71.21% and an F1-score of 70.31%. Although the performance gap is

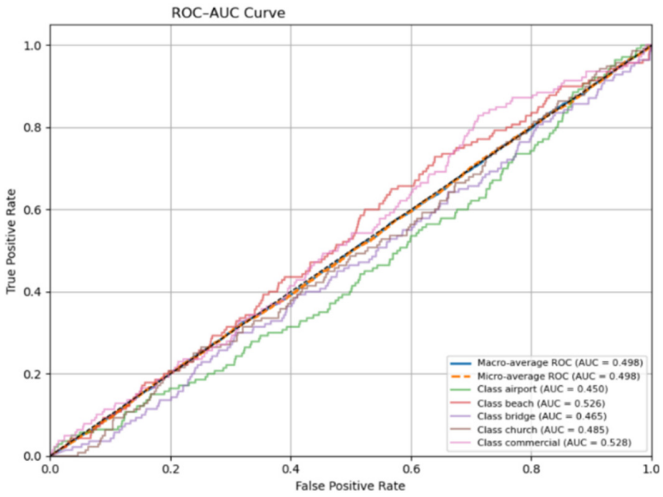
slightly larger than in Scenario 3, ConvNeXt-Tiny still demonstrates competitive robustness under significant domain shift. Again, MobileNetV2 and DenseNet121 showed limited generalization ability, with accuracy values below 55%, highlighting their sensitivity to differences in dataset characteristics.

TABLE V. EVALUATION RESULTS OF SCENARIO 4

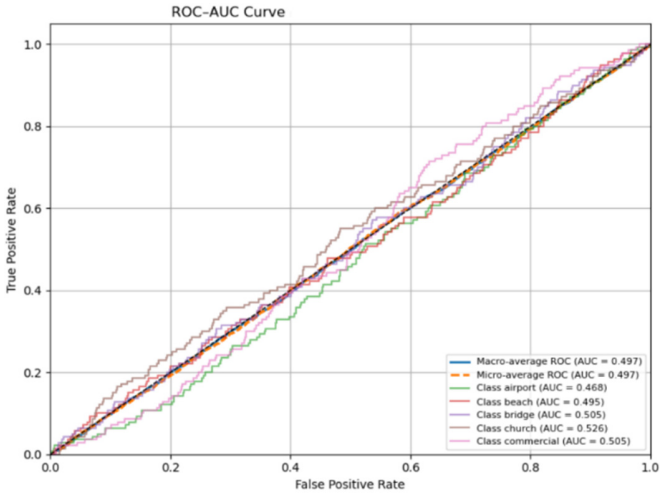
Model	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
ResNet50	72.94	73.87	75.73	72.99
MobileNetV2	39.91	40.89	45.29	37.66
DenseNet121	52.74	53.64	60.04	49.08
ConvNeXt-Tiny	70.31	71.21	72.82	70.56

The performance degradation observed in cross-dataset scenarios can be attributed to domain shift between NWPU-RESISC45 and AID, including differences in spatial resolution, scene composition, and acquisition conditions. NWPU-RESISC45 contains more diverse and fine-grained scene variations, whereas AID exhibits higher inter-class similarity in certain categories. Under these conditions, ConvNeXt-Tiny demonstrates stronger robustness than DenseNet121 and MobileNetV2, suggesting that its modern convolutional design provides more transferable feature representations across datasets and maintains effective discrimination even under domain shifts.

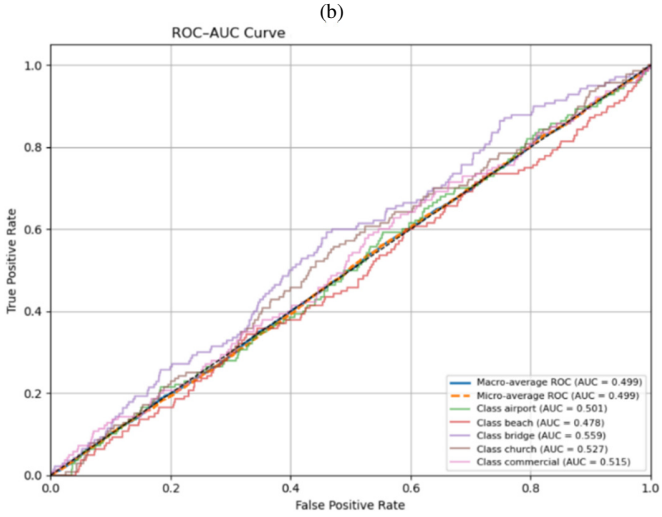
Figure 2 illustrates the ROC-AUC curves of all compared models under Scenario 1, where both training and testing were conducted on the NWPU-RESISC45 dataset. Overall, all models demonstrate AUC values well above the random baseline, indicating effective discriminative capability for RSSC. Among the evaluated models, ConvNeXt-Tiny and ResNet50 exhibit ROC curves that remain consistently closer to the upper-left corner, reflecting strong separability between scene categories and stable classification behavior across different decision thresholds. This observation is consistent with their superior performance in terms of accuracy and F1-score reported in Table II.



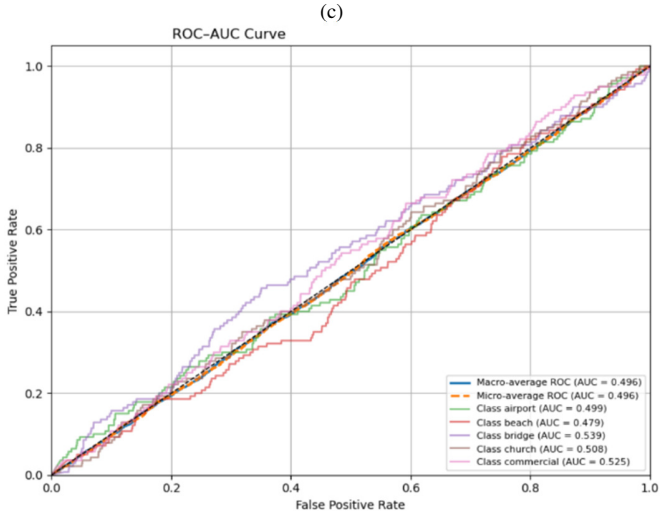
(a)



(b)



(c)



(d)

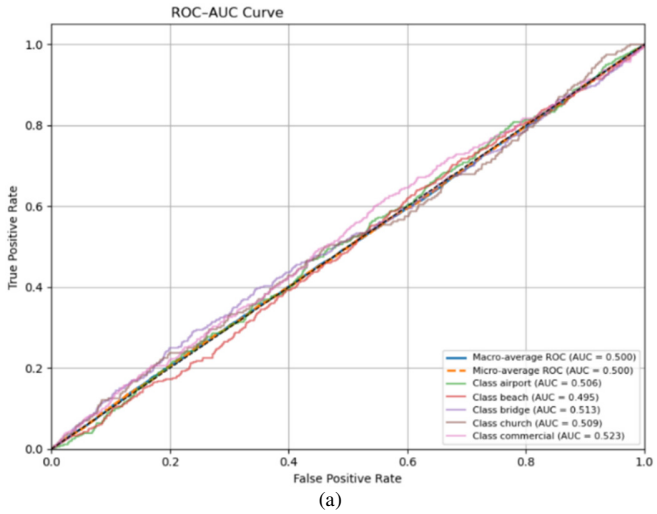
Fig. 2. ROC-AUC curves for Scenario 1: (a) ResNet50, (b) MobileNetV2, (c) DenseNet121, (d) ConvNeXt-Tiny.

In contrast, DenseNet121 and MobileNetV2 show relatively flatter ROC curves, indicating weaker confidence in

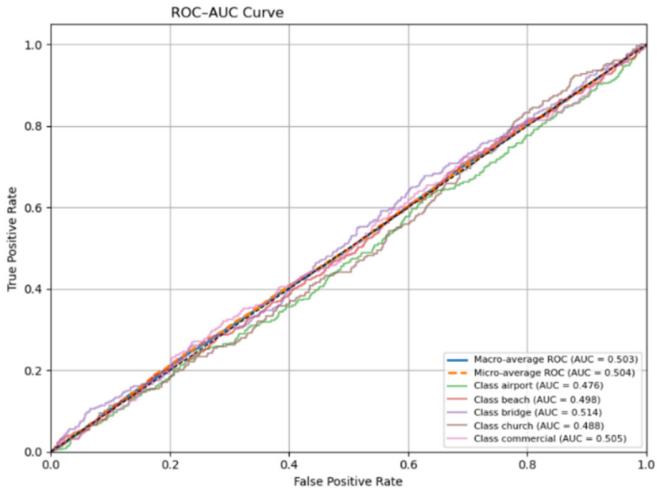
distinguishing between classes. The reduced AUC performance of MobileNetV2 suggests that lightweight architectures may struggle to capture the complex spatial patterns and semantic variability present in the NWPU dataset. Overall, the ROC-AUC analysis under Scenario 1 confirms that ConvNeXt-Tiny achieves a robust and balanced classification performance, comparable to or slightly exceeding that of ResNet50, while maintaining the advantages of a modern convolutional design.

Figure 3 illustrates the ROC-AUC curves of the evaluated models under Scenario 3, where the models were trained on the NWPU-RESISC45 dataset and directly evaluated on the AID dataset without additional fine-tuning. This scenario represents a challenging cross-dataset setting due to domain shift between the two datasets, including differences in spatial resolution, scene composition, and acquisition conditions.

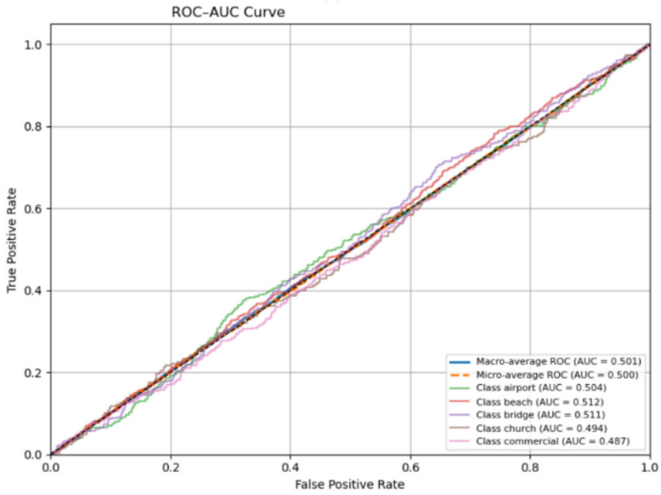
As observed in the figure, ResNet50 and ConvNeXt-Tiny exhibit ROC curves that remain relatively closer to the upper-left corner compared to the other models, indicating stronger discriminative capability under unseen data distributions. Although a general performance degradation is evident compared to in-dataset scenarios, both models maintain stable separation between positive and negative samples. ConvNeXt-Tiny demonstrates a competitive ROC-AUC profile, comparable to that of ResNet50. This observation aligns with the quantitative results reported in Table IV, where ConvNeXt-Tiny achieved an accuracy of 80.31% and an F1-score of 78.63%, closely following ResNet50. The ROC curve suggests that ConvNeXt-Tiny preserves robust feature representations despite the significant domain shift. In contrast, DenseNet121 shows a noticeable decline in ROC-AUC performance, indicating reduced class separability under cross-dataset conditions. MobileNetV2 exhibits the weakest ROC-AUC behavior, with its curve approaching the diagonal line, reflecting limited robustness and sensitivity to dataset variations. Overall, the ROC-AUC analysis for Scenario 3 confirms that ConvNeXt-Tiny maintains strong generalization capability and discriminative power in cross-dataset RSSC, outperforming lightweight architectures and approaching the performance of deeper CNN models.



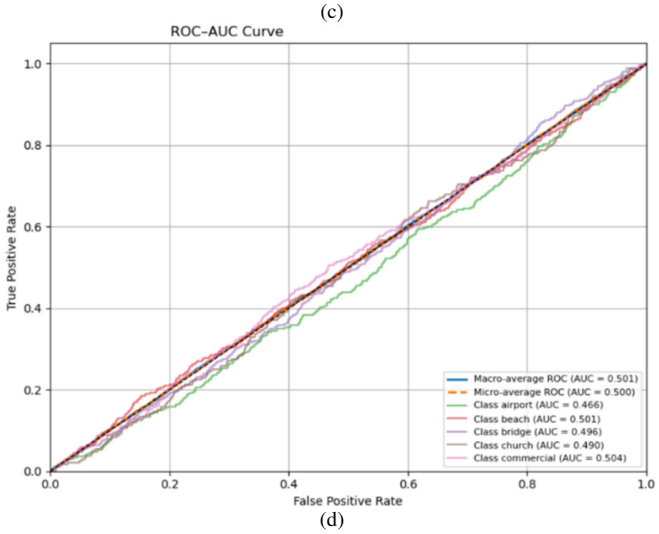
(a)



(b)



(c)



(d)

Fig. 3. ROC-AUC curves for Scenario 3: (a) ResNet50, (b) MobileNetV2, (c) DenseNet121, (d) ConvNeXt-Tiny.

Figures 4 and 5 show the confusion matrices for ResNet50 and ConvNeXt-Tiny under Scenarios 1 and 3. ConvNeXt-Tiny produces fewer misclassifications across several scene

categories compared to lightweight models, particularly in visually similar classes. In cross-dataset evaluation (Scenario 3), an increase in misclassification is observed for all models due to domain shift; however, ConvNeXt-Tiny maintains a more balanced error distribution than DenseNet121 and MobileNetV2, indicating stronger robustness under distribution mismatch. Across all experimental scenarios, ConvNeXt-Tiny consistently demonstrates strong in-dataset performance and competitive cross-dataset generalization, outperforming lightweight architectures and closely matching the performance of deeper CNNs such as ResNet50. These findings highlight the effectiveness of modern convolutional designs for robust RSSC under both controlled and domain-shift conditions.

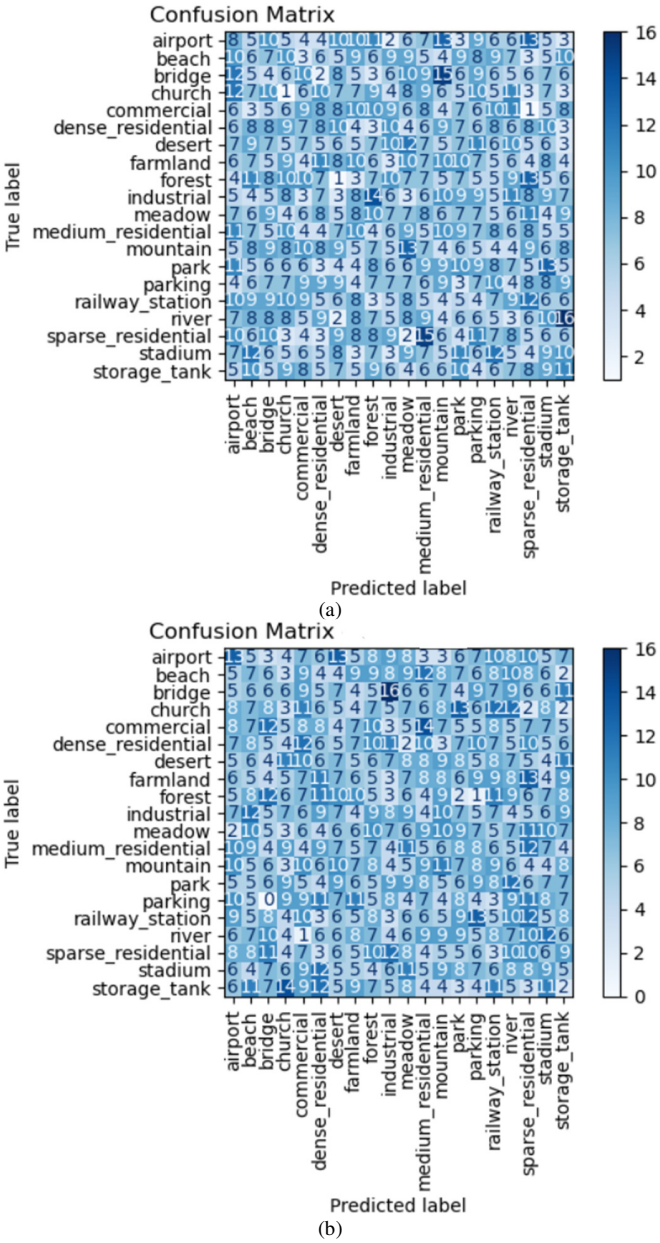


Fig. 4. Confusion matrices for Scenario 1: (a) ResNet50, (b) ConvNeXt-Tiny.

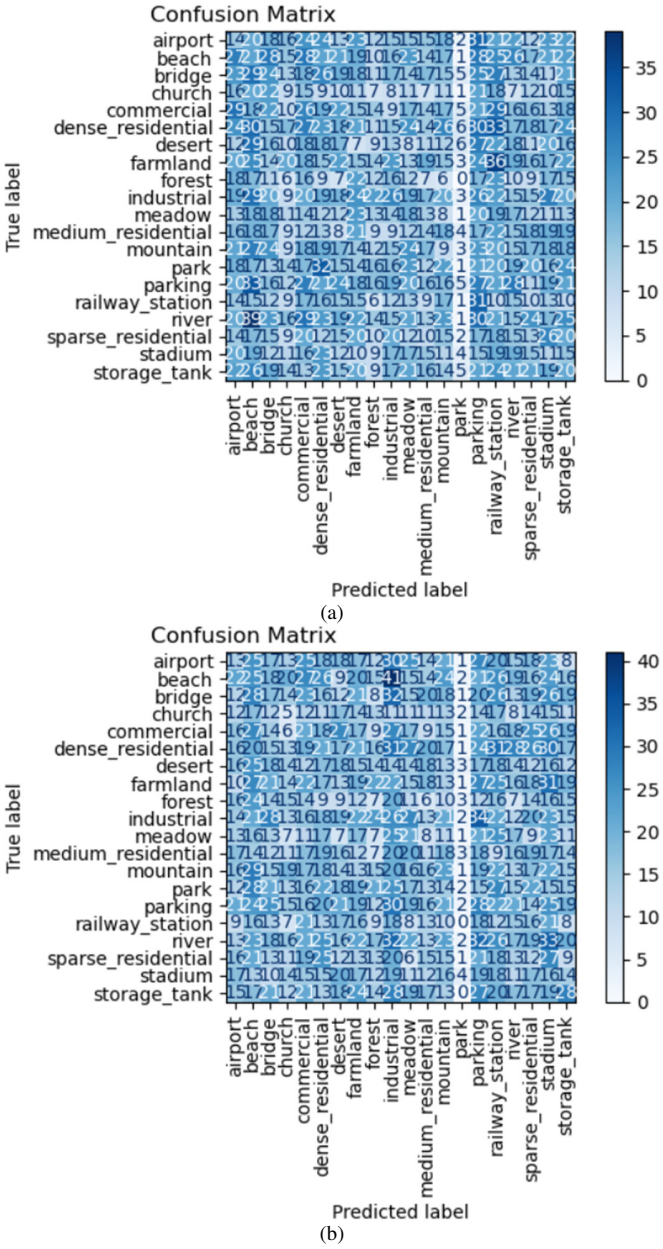


Fig. 5. Confusion matrices for Scenario 3: (a) ResNet50, (b) ConvNeXt-Tiny.

Table VI presents a comparative overview of the proposed method against representative RSSC approaches reported in the literature. The experiments in this study were conducted on the NWPU-RESISC45 and AID datasets, which are widely adopted benchmarks for RSSC. NWPU-RESISC45 consists of 45 scene categories with diverse spatial resolutions and significant intra-class variability, whereas AID contains 30 aerial scene categories characterized by notable inter-class similarity. While some existing methods achieve higher accuracy through architectural enhancements such as attention mechanisms or transformer-based designs, the proposed ConvNeXt-Tiny model demonstrates competitive performance using a pure CNN transfer learning framework. It should be

emphasized that the comparative results in Table VI are not obtained under identical experimental settings. Differences in data splits, augmentation strategies, and training protocols may affect absolute performance values. Therefore, the comparison is intended to provide contextual insight rather than serving as a direct benchmark under strictly controlled conditions.

TABLE VI. COMPARISON WITH EXISTING RSSC METHODS ON BENCHMARK DATASETS

Method	Backbone	Dataset	Accuracy (%)
CNN [9]	Fine-tuned pretrained CNN	NWPU-RESISC45, AID	99.76, 93.89
RS-CLIP [21]	—	NWPU-RESISC45, AID-30	85.76, 87.52
ConvNeXt-Tiny [22]	—	UC-Merced, AID, WHU-RS19	99.00, 95.08, 99.50
Proposed method	ConvNeXt-Tiny	NWPU-RESISC45	90.89
Proposed method	ConvNeXt-Tiny	AID	89.52

VI. CONCLUSIONS

This study presented a comprehensive evaluation of ConvNeXt-Tiny for Remote Sensing Scene Classification (RSSC) using a transfer learning framework and benchmark datasets. By conducting both in-dataset and cross-dataset experiments on NWPU-RESISC45 and AID, this work aimed to assess not only classification accuracy but also the generalization capability of modern Convolutional Neural Networks (CNNs) under domain shift conditions. The experimental results show that ConvNeXt-Tiny consistently achieves strong performance across all evaluation scenarios. In in-dataset experiments, ConvNeXt-Tiny attains accuracy and F1-score values comparable to or slightly higher than ResNet50, demonstrating its effectiveness in capturing discriminative scene representations. In cross-dataset scenarios, although performance degradation is observed for all models due to differences in scene composition, resolution, and acquisition conditions, ConvNeXt-Tiny maintains competitive robustness and closely follows the performance of ResNet50, while significantly outperforming DenseNet121 and MobileNetV2. The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) analysis further supports these findings, indicating stable class separability and reliable decision behavior across varying thresholds.

Overall, the results suggest that ConvNeXt-Tiny combines the advantages of modern convolutional design with strong generalization capability, making it a promising backbone for RSSC tasks. Future work may explore fine-tuning strategies, multi-scale feature learning, or domain adaptation techniques to further enhance cross-dataset performance and extend the applicability of ConvNeXt-based models to more diverse remote sensing scenarios.

In contrast to most existing ConvNeXt-based approaches that rely heavily on data augmentation strategies, attention mechanisms, or hybrid architectural enhancements to boost in-dataset accuracy, this study deliberately focuses on evaluating ConvNeXt-Tiny as a pure convolutional transfer learning backbone under a standard and reproducible supervised learning setting. The primary contribution of this work lies in its emphasis on cross-dataset generalization rather than augmentation-driven performance gains.

By systematically analyzing ConvNeXt-Tiny across different remote sensing datasets without introducing additional architectural complexity, this study provides clearer insights into the robustness and transferability of modern CNN representations under domain shift conditions. These findings are particularly relevant for real-world RSSC systems, where models are often deployed in environments that differ significantly from their training data, and excessive reliance on dataset-specific augmentation techniques may limit reproducibility and generalization. Therefore, this work highlights the practical applicability of ConvNeXt as a reliable and efficient backbone for realistic RSSC scenarios.

ACKNOWLEDGMENT

The authors acknowledge the support o DPPM Kemdiktisaintek and Universitas Islam Lamongan for funding this research.

REFERENCES

[1] A. Sharma, X. Liu, X. Yang, and D. Shi, "A patch-based convolutional neural network for remote sensing image classification," *Neural Networks*, vol. 95, pp. 19–28, Nov. 2017, https://doi.org/10.1016/j.neunet.2017.07.017.

[2] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 197–209, Nov. 2018, https://doi.org/10.1016/j.isprsjprs.2018.01.004.

[3] S. Dutta and M. Das, "Remote sensing scene classification under scarcity of labelled samples—A survey of the state-of-the-arts," *Computers & Geosciences*, vol. 171, Feb. 2023, Art. no. 105295, https://doi.org/10.1016/j.cageo.2022.105295.

[4] A. Thapa, T. Horanont, B. Neupane, and J. Aryal, "Deep Learning for Remote Sensing Image Scene Classification: A Review and Meta-Analysis," *Remote Sensing*, vol. 15, no. 19, Oct. 2023, Art. no. 4804, https://doi.org/10.3390/rs15194804.

[5] H. Li *et al.*, "RSI-CB: A Large-Scale Remote Sensing Image Classification Benchmark Using Crowdsourced Data," *Sensors*, vol. 20, no. 6, Mar. 2020, Art. no. 1594, https://doi.org/10.3390/s20061594.

[6] J. Liang, Y. Deng, and D. Zeng, "A Deep Neural Network Combined CNN and GCN for Remote Sensing Scene Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4325–4338, 2020, https://doi.org/10.1109/JSTARS.2020.3011333.

[7] C.-H. Lin and T.-Y. Wang, "A novel convolutional neural network architecture of multispectral remote sensing images for automatic material classification," *Signal Processing: Image Communication*, vol. 97, Sept. 2021, Art. no. 116329, https://doi.org/10.1016/j.image.2021.116329.

[8] 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. 3735–3756, 2020, https://doi.org/10.1109/JSTARS.2020.3005403.

[9] C. Broni-Bediako, Y. Murata, L. H. B. Mormille, and M. Atsumi, "Searching for CNN Architectures for Remote Sensing Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, Art. no. 4701813, https://doi.org/10.1109/TGRS.2021.3097938.

[10] R. Cao, L. Fang, T. Lu, and N. He, "Self-Attention-Based Deep Feature Fusion for Remote Sensing Scene Classification," *IEEE Geoscience and*

Remote Sensing Letters, vol. 18, no. 1, pp. 43–47, Jan. 2021, https://doi.org/10.1109/LGRS.2020.2968550.

[11] S. Thirumaladevi, K. Veera Swamy, and M. Sailaja, "Remote sensing image scene classification by transfer learning to augment the accuracy," *Measurement: Sensors*, vol. 25, Feb. 2023, Art. no. 100645, https://doi.org/10.1016/j.measen.2022.100645.

[12] A. Ma, Y. Wan, Y. Zhong, J. Wang, and L. Zhang, "SceneNet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 172, pp. 171–188, Feb. 2021, https://doi.org/10.1016/j.isprsjprs.2020.11.025.

[13] J. Shen, B. Cao, C. Zhang, R. Wang, and Q. Wang, "Remote Sensing Scene Classification Based on Attention-Enabled Progressively Searching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, Art. no. 4707513, https://doi.org/10.1109/TGRS.2022.3186588.

[14] S.-B. Chen, Q.-S. Wei, W.-Z. Wang, J. Tang, B. Luo, and Z.-Y. Wang, "Remote Sensing Scene Classification via Multi-Branch Local Attention Network," *IEEE Transactions on Image Processing*, vol. 31, pp. 99–109, 2022, https://doi.org/10.1109/TIP.2021.3127851.

[15] Z. Zhou *et al.*, "NaSC-TG2: Natural Scene Classification With Tiangong-2 Remotely Sensed Imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3228–3242, 2021, https://doi.org/10.1109/JSTARS.2021.3063096.

[16] W. Chen, S. Ouyang, W. Tong, X. Li, X. Zheng, and L. Wang, "GCSANet: A Global Context Spatial Attention Deep Learning Network for Remote Sensing Scene Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1150–1162, 2022, https://doi.org/10.1109/JSTARS.2022.3141826.

[17] J. Zhang, H. Zhao, and J. Li, "TRS: Transformers for Remote Sensing Scene Classification," *Remote Sensing*, vol. 13, no. 20, Oct. 2021, Art. no. 4143, https://doi.org/10.3390/rs13204143.

[18] A. Sivasubramanian, V. Prashanth, T. Hari, V. Sowmya, E. A. Gopalakrishnan, and V. Ravi, "Transformer-based convolutional neural network approach for remote sensing natural scene classification," *Remote Sensing Applications: Society and Environment*, vol. 33, Jan. 2024, Art. no. 101126, https://doi.org/10.1016/j.rsase.2023.101126.

[19] F. Zheng, S. Lin, W. Zhou, and H. Huang, "A Lightweight Dual-Branch Swin Transformer for Remote Sensing Scene Classification," *Remote Sensing*, vol. 15, no. 11, May 2023, Art. no. 2865, https://doi.org/10.3390/rs15112865.

[20] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote-Sensing Scene Classification via Multistage Self-Guided Separation Network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023, Art. no. 5615312, https://doi.org/10.1109/TGRS.2023.3295797.

[21] X. Li, C. Wen, Y. Hu, and N. Zhou, "RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision," *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, Nov. 2023, Art. no. 103497, https://doi.org/10.1016/j.jag.2023.103497.

[22] R. A. Perdana, A. M. Arimurthy, and Risnandar, "Remote Sensing Scene Classification using ConvNeXt-Tiny Model with Attention Mechanism and Label Smoothing," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 3, pp. 389–400, June 2024, https://doi.org/10.29207/resti.v8i3.5731.

[23] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, https://doi.org/10.1109/JPROC.2017.2675998.

[24] Z. He, G. Li, Z. Wang, G. He, H. Yan, and R. Wang, "Deep Ensemble Remote Sensing Scene Classification via Category Distribution Association," *Remote Sensing*, vol. 16, no. 21, Nov. 2024, Art. no. 4084, https://doi.org/10.3390/rs16214084.

[25] Z. Li *et al.*, "A Hierarchical Graph-Enhanced Transformer Network for Remote Sensing Scene Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 20315–20330, 2024, https://doi.org/10.1109/JSTARS.2024.3491335.

[26] G.-S. Xia *et al.*, "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, July 2017, https://doi.org/10.1109/TGRS.2017.2685945.

[27] W. Wang, Y. Sun, J. Li, and X. Wang, "Frequency and spatial based multi-layer context network (FSCNet) for remote sensing scene classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 128, Apr. 2024, Art. no. 103781, https://doi.org/10.1016/j.jag.2024.103781.

[28] A. Khaldi and J. A. Jumayel, "Intelligent Remote Sensing Scene Classification Model for On-Board Training of Resource-Constrained Devices," *Journal of Cognitive Human-Computer Interaction*, vol. 9, no. 1, pp. 1–19, Jan. 2025, https://doi.org/10.54216/JCHCI.090101.

[29] R. Zhu, S. Ma, J. Lian, L. He, and S. Mei, "Generating Adversarial Examples Against Remote Sensing Scene Classification via Feature Approximation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 10174–10187, 2024, https://doi.org/10.1109/JSTARS.2024.3399780.

[30] J. Feng, H. Luo, and Z. Gu, "Improving semi-supervised remote sensing scene classification via Multilevel Feature Fusion and pseudo-labeling," *International Journal of Applied Earth Observation and Geoinformation*, vol. 136, Feb. 2025, Art. no. 104335, https://doi.org/10.1016/j.jag.2024.104335.

[31] G. Cheng, J. Han, and X. Lu, "NWPU-RESISC45: Remote sensing image scene classification dataset." OneDrive. [Online]. Available: https://onedrive.live.com/?redeem=aHR0cHM6Ly8xZHJ2Lm1zL3UvcyFBbWdLWXpBUkJsNWNhM0hOYUhJbHpwX0lYanM&cid=5C5E061130630A68&id=5C5E061130630A68%21107&parId=5C5E061130630A68%21112&o=OneUp.

[32] G.-S. Xia *et al.*, "AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification." Github. [Online]. Available: https://captain-whu.github.io/AID/.

AUTHORS PROFILE

Nur Nafiiyah received her Bachelor of Informatics Engineering from Universitas Islam Lamongan, Indonesia (2005-2009), and her Master of Information Technology from the Sekolah Tinggi Teknik Surabaya, Indonesia (2011-2013). She holds a Ph.D. degree in Computer Science from the Department of Informatics, Institut Teknologi Sepuluh Nopember (2019-2023). She is currently interested in artificial intelligence, deep learning, and computer vision. She can be contacted at email: mynaff@unisla.ac.id.

Agus Harjoko received the Bachelor's degree in Electronics and Instrumentation from Universitas Gadjah Mada, Indonesia, in 1986, the Master's degree in Computer Science from the University of New Brunswick, Canada, in 1990, and the Ph.D. degree in Computer Science from the University of New Brunswick, Canada, in 1996. He is currently a Professor in the Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia, where he has held the position since 2023. His research interests include computer vision, pattern recognition, and instrumentation.

Achmad Nizar Hidayanto holds a Bachelor of Computer Science (S.Kom.), a Master of Computer Science (M.Kom.), and a Ph.D. in information system from Faculty of Computer Science, University of Indonesia. He is currently a Professor in the Department of Computer Science, University of Indonesia, Indonesia. His research areas of interest include information technology, information system, business intelligence, technology adoption, and electronic commerce. He can be contacted at email: nizar@cs.ui.ac.id.