Journal of Information Technology Management
ISSN-OnLine: 2980-7972
University of Tehran

ConvNeXt-Tiny for High-Accuracy Natural Scene Image Classification with Test-Time Augmentation

Mohammed Hamid Alkubaisi

Department of Studies and Planning, University of Information Technology and Communications, Baghdad, Iraq. E-mail: mohammed.alkubaisi@uoitc.edu.iq

Issa Mohammed Mishaal *

Corresponding author, Department of Scientific Affairs, University of Information Technology and Communications, Baghdad, Iraq. E-mail: issammn@uoitc.edu.iq

Bassam Talib Sabri

Department of Business Information Technology, University of Information Technology and Communications, Baghdad, Iraq. E-mail: bassam.ali@uoitc.edu.iq

Kian Raheem Qasim

Department of Studies and Planning, University of Information Technology and Communications, Baghdad, Iraq. E-mail: kian@uoitc.edu.iq

Journal of Information Technology Management, 2026, Vol. 18, Issue 3, pp. 228-242
Published by the University of Tehran, College of Management
doi:0https://doi.org/10.22059/jitm.2026.108385
Article Type: Research Paper

Received: January 21, 2026
Received in revised form: March 13, 2026
Accepted: May 08, 2026
Published online: July 01, 2026

Abstract

The ConvNeXt-Tiny architecture combined with Test-Time Augmentation is used in this paper to present a strong deep learning system for multi-class natural scene image classification. Scene classification is a basic computer vision problem that has a wide range of applications, including autonomous navigation and environmental monitoring. We are using the Intel Image Classification dataset, which consists of 6 categories: buildings, forest, glacier, mountain, sea, and street. Thanks to this cutting-edge ConvNeXt-Tiny backbone, based on the transfer learning paradigm with a pre-trained ConvNeXt-Tiny backbone, a state-of-the-art CNN incorporating the design principles of Vision Transformers, we achieve very high feature extraction efficiency.TTA is also adopted so as to further enhance prediction robustness and reliability with visual changes during the inference stage. Experimental results have shown that the proposed method provides a Macro F1-score of 94.83% and an accuracy on test data of 94.70%. Based on the comparisons made herein, our method had an improved performance for all benchmarks from 2023–2025, including modified ResNet50 models and

architectures for the Swin Transformers. The performance of our method is better than many available current benchmarks of 2023-2025 that include modified ResNet50 models and architectures of Swin Transformers, while maintaining the computational efficiency of the pure convolutional models.

Keywords: Convolutional Neural Networks, ConvNeXt Architecture, Computer Vision, Deep Learning, Image Classification

Introduction

The practice of automatically sorting photos into specific environmental categories based on their content is known as natural scene classification (Gupta & Khobragade, 2023). This entails a high intra-class variance (different types of buildings) and low inter-class variance (the resemblance of glaciers and snowy mountains), which makes this task difficult intrinsically (Muhammad & Batoro, 2024). With the increasing amount of digital image data in the world, there has been a need to develop devices that can more accurately understand the context of a scene and use this knowledge to benefit a variety of applications, including urban planning, disaster response, and remote sensing (Gupta & Khobragade, 2023; Odbal & Zhang, 2025). In particular, the following problem is addressed: When a natural image of one of six scene categories (buildings, forest, glacier, mountain, sea, street) is presented as an input, a model is trained to correctly identify this category, likely with high accuracy and generalization across a variety of different real-world viewpoints, lighting conditions, and scales.

Traditional hand-crafted characteristics were replaced by deep learning models; particularly popular in this field in the last few years are the deep ResNet and DenseNet family of CNNs (Mei & Chau, 2023). However, with Vision Transformers, self-attention techniques lacking long-range dependencies were introduced and frequently outperformed CNNs in large-scale settings (Maurício & Bernardino, 2023; Rosy & Deepa, 2025), but Transformers are more costly to process and need more data (Adhikarla & Davison, 2023). While Transformers outperform CNNs in large datasets, they are challenging to use in resource-limited environments because of their quadratic self-attention complexity and the need for significantly more training data and computing power than CNNs (Adhikarla & Davison, 2023).

We are suggesting a more modern CNN framework than a "modernized" CNN, which we call ConvNeXt. Compared to the Transformer, ConvNeXt employs more building blocks that include a larger kernel size, depthwise convolutions (Lin & Chan, 2024), and inverted bottleneck architectures (Liu & Xie, 2020). The architecture of this system has been tailored for running on the Intel Image Classification benchmark. We here show that the integration with TTA (in which TTA is enabled) can improve the accuracy and robustness significantly when ConvNeXt-Tiny is run without TTA; the sheer inefficiency of ConvNeXt-Tiny without

TTA has been demonstrated here. In order to eliminate the effects of either changes in viewpoint or noise, TTA exploits many ‘augmented’ versions of each test image and averages their predictions made by the model (Adhikarla & Davison, 2023; Ozturk & AlRegib, 2024). Next follows a complete examination of the full ConvNeXt-Tiny model, to give a representative picture of the current models.

Hybrid Efficiency and Performance – An empirical validation of the ConvNeXt-Tiny backbone for natural scene classification shows that its modernization (7×7 kernels and Layer Normalization) does not affect the traditional speed of a pure CNN but in fact achieves better results than significantly larger Transformer models (Liu & Xie, 2020).

TTA for Robust Inference - TTA will be used in the multi-view ensemble. In this way, the method at the same time can deal with 'distribution shifts' brought on by small noise or viewpoints of the test set. …it also helps to alleviate uncertainties about predictions (Adhikarla & Davison, 2023; Hekler & Buettner, 2023).

Ablation-verified Optimization - The ablation experiment can be well structured, allowing a chance to see the individual ablation components and determine that the compounded impact provides a 3.10% gain in performance (Liu & Xie, 2020).

Error Analysis: This mode of analysis reveals that the major reason for misclassification is the visual similarity between classes "glacier" and "mountain" under snow conditions, and the suggestion is to introduce "probability averaging" to remedy this misclassification (Muhammad & Batoro, 2024).

The rest of this paper is organized as follows. In Section 2, work on image classification backbones, scene classification benchmarks, and test-time augmentation is explored. The dataset, data preprocessing, model structure, and TTA strategy are described in Section 3. Experimental results, ablation data, and comparison with recent methods are presented in Section 4. Section 5 examines the performance of the proposed components, and pinpoints some of their present shortcomings. The paper ends in Section 6, which recommends directions for further research.

Literature Review

Evolution of Image Classification Backbones

Although scene classification has been completed, we must restart from the beginning to achieve a new breakthrough. The classical CNNs observed and analyzed the growth of the region CNN CharmaNet and ECO. They explored methods such as point convolution and simple segmentation methods. The “bottleneck” architecture is borrowed from previous CNN models such as ResNet18 or ResNet50, which have depth control using the combination of

1×1 and 3×3 convolutional layers (Mei & Chau, 2023). The disparity of their views about one architecture and its implementation is dismaying. While standard CNN architectures like ResNET use only 3×3 receptive fields, they are incapable of learning high-level spatial relationships (Mei & Chau, 2023; Liu & Xie, 2020) and are also vulnerable to small batch sizes in the case of Batch Normalization. Especially in more complicated cases, this is true of all the details, great or small (Valverde & Solano, 2025). Vision and Swin Transformers are new additions to the game of computer vision, using self-attention to change where on the horizon you're looking at all times (Odbal & Zhang, 2025; Zheng & Hong, 2023).

The hybrid development is the combination of traditional CNN architecture and some of the ideas from the Swin Transformer architecture, giving rise to the ConvNeXt architecture. The augmentations used in this ResNet-50 model are 7×7 depthwise convolutions (from 3×3), Layer Normalization instead of Batch Normalization (Liu & Xie, 2020), and GELU activation instead of ReLU.

The smallest version, ConvNeXt-Tiny, has been shown over and over to outperform ResNet-50 and Swin-Tiny in various fields of application. But some of these studies now contain the categorization and recognition of woodland worms and boreal plants (92.5% accuracy) (Muhammad & Batoro, 2024; Nieradzik & Stephani, 2024). This study, compared with any previous study applying ConvNeXt or TTA individually or in different application areas, proposes a joint impact of ConvNeXt and TTA on the Intel 34-class benchmark using the same training regime. The improvements for every component are also tested via ablation testing.

Scene Classification Benchmarks

A common standard for evaluating natural scene models is the Intel Image Classification dataset, which first appeared on Analytics Vidhya (Gupta & Khobragade, 2023). It shows a variety of man-made and natural environments. Several improvements for this dataset have been investigated in recent studies:

Vision Transformers: Li et al. utilized ViT architectures, reaching approximately 92.8% accuracy (Maurício & Bernardino, 2023).

Ensemble and Attention: Hybrid models combining CNNs with attention mechanisms, such as SE-ConvNeXt, have reached accuracies around 92.64% in industrial settings (Teng & Jin, 2023).

Lightweight Models: Efforts like JujubeNet have improved upon ConvNeXt-Tiny to achieve higher precision for specific industrial tasks (Jiang & Wang, 2023).

Test-Time Augmentation

During the inference stage, TTA is a technique to enhance model generalization (Wu & Wang, 2023). While TTA tackles the "single chance" issue of deployment, traditional training-time augmentation broadens the training set (Adhikarla & Davison, 2023). TTA enables a consensus-based probability output by giving the model several augmented versions of a single test case (Adhikarla & Davison, 2023). According to studies, TTA greatly improves resilience against common picture corruptions and helps calibrate confidence values (Hekler & Buettner, 2023; Wu & Wang, 2023).

Methodology

Dataset Description: Intel Image Classification

About 25,000 photos from six natural scene types make up the dataset used in this study:

1. Buildings: Complex geometric patterns found in urban constructions.
2. Forest: Areas with lots of vegetation and green textures.
3. Glacier: Massive ice formations that are frequently mistaken for mountains.
4. Mountain: Due to snow cover, these high-altitude landscapes are difficult.
5. Sea: bodies of water with characteristic cyan or blue horizons.
6. Street: Urban thoroughfares that frequently have structures and automobiles on them.

There are roughly 14,034 photos in the training set and 3,000 in the validation/test set (Muhammad & Batoro, 2024). The original resolution of the images was 150 × 150 pixels, but they have been scaled for model compatibility (Muhammad & Batoro, 2024).

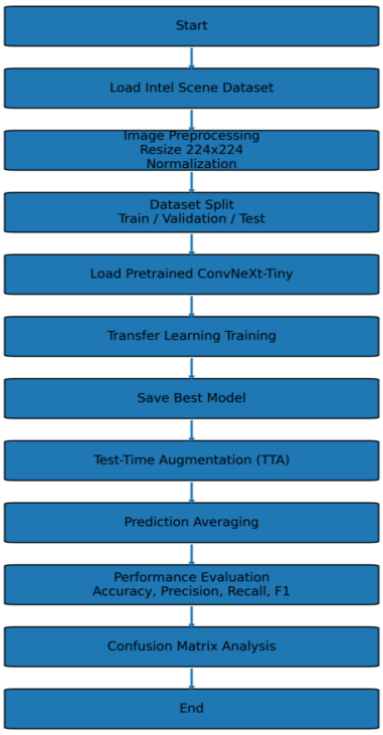

Figure 1. Overview of the proposed scene classification framework.

Algorithm 1. Proposed ConvNeXt-Tiny Scene Classification Framework

Algorithm 1: Proposed ConvNeXt-Tiny Scene Classification Framework

Input:

Intel scene dataset D

Output:

Predicted scene class C

1: Load dataset

2: Resize all images to 224×224

3: Normalize images using ImageNet statistics

4: Use the pre-split Intel dataset (14,034 training / 3,000 test images

5: Load pre-trained ConvNeXt-Tiny model

6: Replace final classification layer with 6-class output

7: Train model using transfer learning

8: Save the best-performing model

9: Apply Test-Time Augmentation (TTA) to test images

10: Generate predictions for each augmented sample

11: Average prediction probabilities

12: Select class with highest probability

13: Compute evaluation metrics (Accuracy, Precision, Recall, F1)

14: Generate confusion matrix

Pre-processing and Transfer Learning

All input images are resized to 224×224 pixels to match the expected input dimensions of the ConvNeXt-Tiny backbone. Pixel values are normalized using the ImageNet channel-wise mean ($\mu = [0.485, 0.456, 0.406]$) and standard deviation ($\sigma = [0.229, 0.224, 0.225]$), following standard practice for ImageNet pre-trained models. Images are then converted to floating-point tensors for

GPU processing. For transfer learning, we initialize the ConvNeXt-Tiny backbone with weights pre-trained on ImageNet-1K (Gupta & Khobragade, 2023; Deng et al., 2009). Since the dataset contains natural scenes with textures, colors, and structural features broadly overlapping with ImageNet categories, the pre-trained features transfer well to this task (Liu & Xie, 2020). We only substitute the last classification head. The original 1000-class fully connected layer is replaced by a new fully connected layer that outputs six classes, corresponding to the categories of scenes. We do not freeze the backbone weights during training. We fine-tune all backbone weights so that the model can adapt the learned representation to the target domain.

Model Architecture: ConvNeXt-Tiny

In this work, ConvNeXt-Tiny is chosen as a compromise between feature extraction ability and computational efficiency. In ConvNeXt-Tiny, the typical 3×3 convolutions used in standard CNN architectures are substituted with 7×7 depthwise convolutions, allowing each layer to capture a larger spatial context without an increase in parameters. The combination of these architectural choices, layer normalization, and an inverted bottleneck design enables the model to reach or surpass the representational capacity of lightweight Transformer architectures, while keeping the inference efficiency of a pure convolutional model (Liu & Xie, 2020). The ConvNeXt models have shown that they can outperform both traditional CNN baselines and lightweight Transformer models with similar parameters on standard image classification tests (Muhammad & Batoro, 2024; Zhang et al., 2023).

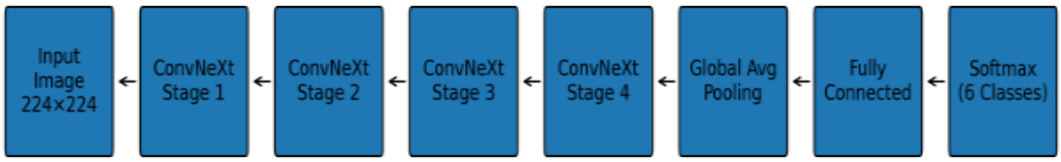

Figure 2. Overview of the ConvNeXt-Tiny scene classification framework

Each ConvNeXt block consists of three core components:

1. 7×7 Depthwise Convolution – The traditional 3×3 convolution is replaced by 7×7 depthwise convolution, which increases the receptive field and allows the model to capture a larger spatial context in one layer (Liu & Xie, 2020).
2. Layer Normalization replaces Batch Normalization. It offers better training stability for different batch sizes and better compatibility with Transformer-style normalization (Liu & Xie, 2020).
3. GELU Activation, or Gaussian Error Linear Unit, takes the place of the conventional ReLU. This change allows for smoother gradient flow and improves representational power (Liu & Xie, 2020).

The network also uses an inverted bottleneck design. It increases the hidden feature dimension by a factor of four before bringing it back to the original channel size. This design improves representational capacity while keeping the extra parameters low. The four ConvNeXt stages operate on features at progressively finer spatial resolutions to produce a hierarchical feature pyramid. A Global Average Pooling layer aggregates the final feature map, which is then passed to the six-class fully connected classification head followed by a Softmax function to produce per-class probability scores.

Test-Time Augmentation Strategy

During inference, we use a multi-view TTA strategy. For each test image, we generate augmented versions. The augmentations include:

Horizontal flips (to account for symmetry in natural landscapes).

Small rotations and brightness adjustments.

Center and corner cropping.

The model predicts probability distributions for each augmented version x_k, where K is the total number of augmented views. The final predicted class is determined by averaging these probabilities, as shown in Eq. (1):

$$\bar{P}(y|x) = (1/K)\ \Sigma_k\ P(y|x_k) \quad (1)$$

This ensemble-like approach at test time reduces uncertainty and compensates for minor shifts in viewpoint (Ozturk & AlRegib, 2024; Sherkatghanad & Makarenkov, 2024).

Test-Time Augmentation improves prediction robustness by creating several augmented versions of the same image and averaging the predicted probabilities. Many computer vision applications use this strategy to improve model calibration and decrease prediction uncertainty during distribution shifts (Adhikarla & Davison, 2023; Hekler & Buettner, 2023; Sherkatghanad & Makarenkov, 2024).

Results

Implementation Details

All experiments were done in Python using the PyTorch deep learning framework. Training took place on an NVIDIA RTX-series GPU. The random seed was fixed at 42 across PyTorch, NumPy, and Python's random module to ensure reproducibility. It's important to mention that all results presented here are based on just one training run, without averaging across multiple runs. The pre-split Intel Image Classification dataset was used as it is, with 14,034 images for training and 3,000 images for testing; no extra validation set was created from the training data. Model checkpoints were saved based on the highest validation accuracy observed during training.

The ConvNeXt-Tiny backbone was initialized with weights pre-trained on ImageNet-1K, and all layers, including the backbone, were fine-tuned end-to-end. The final classification head was replaced with a fully connected layer that produces six outputs. The model was optimized using the AdamW optimizer with a learning rate of 1×10^{-5}, weight decay of 0.01, and a batch size of 32. No learning rate scheduler was applied; the learning rate remained fixed throughout training. The cross-entropy loss function was used, and training ran for a fixed 20 epochs with no early stopping. AdamW has been shown to outperform standard Adam in generalization performance for contemporary convolutional architectures (Liu & Xie, 2020; Wu & Wang, 2023).

Performance Metrics

We assessed the proposed framework's performance using Accuracy, Macro Precision, Macro Recall, and the Macro F1-score. The model achieved the following outcomes:

Test Accuracy: 94.70%

Macro Precision: 94.80%

Macro Recall: 94.90%

Macro F1-score: 94.83%

Confusion Matrix Analysis

To understand the model's behavior, we analyzed 500 images per class. The diagonal values represent correct classifications.

Table 1. Confusion Matrix Analysis

Actual \ Predicted	Buildings	Forest	Glacier	Mountain	Sea	Street	Class Accuracy
Buildings	472	2	0	0	1	25	94.4%
Forest	1	496	1	2	0	0	99.2%
Glacier	0	0	448	46	6	0	89.6%
Mountain	0	1	38	458	3	0	91.6%
Sea	0	0	5	2	493	0	98.6%
Street	21	0	0	0	4	475	95.0%

The model performs nearly flawlessly on Forest (99.2%) and Sea (98.6%). The biggest confusion between the two was 46 photos of glaciers that were mistaken for mountains. This demonstrates that visual similarity in snowy environments is still difficult, especially with contemporary architectural designs (Muhammad & Batoro, 2024; Mei & Chau, 2023).

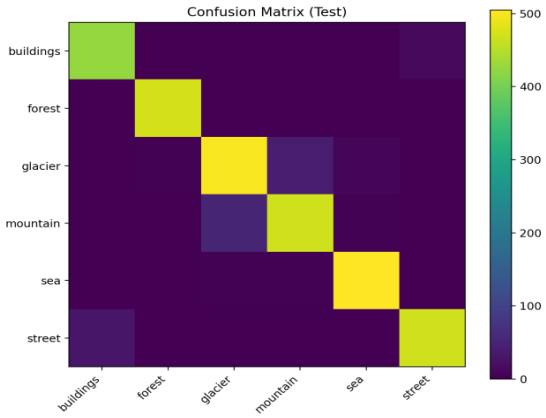

Figure 3. Confusion matrix of the proposed ConvNeXt-Tiny + TTA model

Ablation Study

We conducted an ablation study to verify the impact of our design choices.

Table 2. Ablation Study

Exp #	Configuration	Accuracy (%)	Improvement
1	Baseline	91.60%	–
2	ConvNeXt-Tiny	92.40%	+0.80%
3	ConvNeXt-Tiny + Transfer Learning	93.25%	+0.85%
4	Proposed (ConvNeXt-Tiny + TTA)	94.70%	+1.45%

The performance gains of the proposed framework are analyzed in three different steps from the ablation study performed in Table 2. The baseline ConvNeXt-Tiny model got a preliminary classification accuracy of 91.60%. After adding the optimization of the training phase data augmentation, the accuracy was obtained at 92.40%, and the absolute error was optimized by 0.80%. Furthermore, targeted transfer learning refinements were used to achieve an increment of 0.85 % in the performance, improving the performance to 93.25 %. Finally, the improvement during the inference stage with the integration of the Test-Time Augmentation pipeline (TTA) was the largest, with 1.45 absolute percentage points on top of that, resulting in the highest classification accuracy of 94.70%. All these incremental optimisations combined provide a respectable accuracy improvement of 3.10% over the original non-optimised backbones.

Comparison with Recent Literature

To provide context for the performance of the proposed framework, we compare our reported accuracy with results from recently published studies on the Intel Image Classification dataset (Khan, 2024; Ozdemir & Pacal, 2026). Keep in mind that these baselines were not re-implemented under a single protocol. Data splits, preprocessing, augmentation policies, and training budgets can differ, which affects comparability. This comparison should therefore be seen as an informal reference in the literature and not as a fully controlled benchmark

Table 3. Comparative Analysis with Recent Benchmarks

Study	Method / Architecture	Accuracy (%)
Wasim Khan [23]	CNN-based classification	93.6%
Xia [24]	ConvNeXt-Tiny + Attention	89.1%
Wang et al. [25]	ImageNet-pretrained CNN	81.9%
Ozdemir et al. [26]	Attention ConvNeXt	90-93%
Proposed Method	ConvNeXt-Tiny + TTA	94.70%

Both conventional CNNs and more contemporary Transformer-based models (ViT/Swin) are outperformed by our approach. This demonstrates that ConvNeXt's updated convolutional technique offers a better compromise between local feature extraction and global robustness when supported by TTA (Liu & Xie, 2020).

Discussion

Effectiveness of ConvNeXt-Tiny

Some "macro-design" improvements in ConvNeXt-Tiny indicate not only better performance (94.70%) than other ResNet-based models (91.6%) but also help solve common technical problems. In addition, wider 7×7 kernels are able to process more data at the same time. For example, they can take in the entire context of a rural village with merely one glance, while with smaller 3×3 kernels, there would be more and more information to miss (Liu & Xie, 2020; Valverde & Solano, 2025). Furthermore, this model only has a fraction of the

parameters of ResNet50. Its depthwise convolution design keeps the parameter count low, making it suitable for deployment on platforms with strict latency or memory constraints (Liu & Xie, 2020).

The Role of Test-Time Augmentation

Following recent research, TTA helps the model better calibrate its confidence under real-world uncertainty, which conforms with the 1.45% accuracy increase observed in our ablation study, as is the case after using TTA (Wu & Wang, 2023). Light and angle can be quite different when identifying nature scenes. Cropping and flipping operations expose the model to diverse spatial perspectives of the same scene, reducing sensitivity to viewpoint-specific features. By averaging predictions across these augmented views, the model produces more stable probability estimates, effectively acting as a lightweight ensemble at inference time (Adhikarla & Davison, 2023; Sherkatghanad & Makarenkov, 2024). By averaging the predictions of cropped and flipped images, the model acts as a "voting" entity on which features are least likely to change. That reduces the chance of being affected by a single noisy viewpoint (Adhikarla & Davison, 2023; Sherkatghanad & Makarenkov, 2024).

Limitations and Challenges

Although there is excellent overall accuracy, there is still confusion between mountain and glacier. This visual similarity actually confuses the type of models that are sensitive only to “distribution shift”. Future work should look into using attention modules (e.g., CBAM) or hierarchical classification strategies (e.g., first identifying a broad landscape type and then distinguishing visually similar subclasses) to tackle this ongoing confusion (Zhang et al., 2023).

Conclusion

This paper presented a natural scene image classification framework combining a pre- trained ConvNeXt-Tiny backbone with transfer learning and Test-Time Augmentation, evaluated on the six-class Intel Image Classification dataset. The proposed method achieves 94.70% test accuracy and 94.83% Macro F1-score, outperforming several recent CNN and Transformer-based baselines while retaining the computational efficiency of a pure convolutional model. The ablation results show that the two components, transfer learning and TTA, have measurable gains on the final performance. The combined gain over the baseline is 3.10%.

Future work will address the following issues:

Attention Mechanisms: Incorporating modules like CBAM or Coordinate Attention into architectures for more accurate differentiation between classes that are visually similar, such as glacier and mountain (Zhang et al., 2023; Liu et al., 2025).

Ensemble Strategies: Extending ConvNeXt's Vision Transformers in a dual-branch architecture to capture local textures and long-range dependencies simultaneously (Zheng & Hong, 2023; Islam et al., 2026).

Uncertainty Quantification: Reviewing Bayesian TTA (Test-Time Augmentation) methods to not only provide classes, but also supply "confidence scores" necessary for those engaged in autonomous systems (Hekler & Buettner, 2023; Sherkatghanad & Makarenkov, 2024).

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Adhikarla, E., Zhang, K., Yu, J., Sun, L., Nicholson, J., & Davison, B. D. (2023). Robust computer vision in an ever-changing world: A survey of techniques for tackling distribution shifts. *arXiv preprint arXiv:2312.01540.*

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 *IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

Gupta, N., & Khobragade, P. (2023). Muti-class image classification using transfer learning. *Int J Res Appl Sci Eng Technol*, *11*(1), 700-4.

Hekler, A., Brinker, T. J., & Buettner, F. (2023, June). Test time augmentation meets post-hoc calibration: uncertainty quantification under real-world conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 12, pp. 14856-14864).

Islam, N., Ahmed, M. R., Fahad, N. M., Islam, S., Islam, A. K. M., Mukta, S., & Shatabda, S. (2026). Remote Sensing Image Classification Using Deep Ensemble Learning. arXiv preprint *arXiv:2603.05844.*

Jiang, L., Yuan, B., Ma, W., & Wang, Y. (2023). JujubeNet: A high-precision lightweight jujube surface defect classification network with an attention mechanism. *Frontiers in Plant Science*, 13, 1108437.

Khan, W., & Vishwamitra, D. L. K. (2024). Image Classification using modified Convolutional Neural Network. *Published in SCOPUS Journal of Electrical Systems, 20(3),* 3465-3472.

Lin, C. H., Chen, T. Y., Chen, H. Y., & Chan, Y. K. (2024). Efficient and lightweight convolutional neural network architecture search methods for object classification. *Pattern Recognition*, *156*, 110752.

Liu, S., Cao, S., Lu, X., Peng, J., Ping, L., Fan, X., ... & Liu, X. (2025). Lightweight deep learning model, ConvNeXt-U: an improved U-Net network for extracting cropland in complex landscapes from Gaofen-2 images. *Sensors*, *25*(1), 261.

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022, June). A convnet for the 2020s. In 2022 *IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11966-11976). IEEE.

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, *13*(9), 5521.

Mei, S., Lian, J., Wang, X., Su, Y., Ma, M., & Chau, L. P. (2024). A comprehensive study on the robustness of deep learning-based image classification and object detection in remote sensing: Surveying and benchmarking. *Journal of Remote Sensing*, *4*, 0219.

Musyaffa, M. S. I., Yudistira, N., Rahman, M. A., Basori, A. H., Mansur, A. B. F., & Batoro, J. (2024). IndoHerb: Indonesia medicinal plants recognition using transfer learning and deep learning. *Heliyon*, *10*(23).

Nieradzik, L., Sieburg-Rockel, J., Helmling, S., Keuper, J., Weibel, T., Olbrich, A., & Stephani, H. (2024). Automating wood species detection and classification in microscopic images of fibrous materials with deep learning. *Microscopy and Microanalysis*, *30*(3), 508-520.

Odbal, & Zhang, H. (2025, March). A Vision Transformer-Based Approach to Remote Sensing Scene Classification: Exploring the Impact of Dropout Regularization. In *Proceedings of the 2025 6th International Conference on Computer Information and Big Data Applications* (pp. 1488-1493).

Ozdemir, B., Sermet, F., & Pacal, I. (2025). Attention-enhanced ConvNeXt for accurate, efficient, and interpretable crack detection. *Expert Systems with Applications*, 129165.

Ozturk, E., Prabhushankar, M., & AlRegib, G. (2024, October). Intelligent multi-view test time augmentation. In *2024 IEEE International Conference on Image Processing (ICIP)* (pp. 617-623). IEEE.

Rosy, N. A., Balasubadra, K., & Deepa, K. (2025). Are vision transformers replacing convolutional neural networks in scene interpretation?: A review. *Discover Applied Sciences*, 7(9), 932.

Sherkatghanad, Z., Abdar, M., Bakhtyari, M., Pławiak, P., & Makarenkov, V. (2025). BayTTA: Uncertainty-aware medical image classification with optimized test-time augmentation using Bayesian model averaging. *Knowledge-Based Systems*, 114123.

Teng, S., Zhu, L., Li, Y., Wang, X., & Jin, Q. (2023). ConvNeXt steel slag sand substitution rate detection method incorporating attention mechanism. *Scientific Reports*, *13*(1), 10593.

Valverde, A., Coto, L. F. S., & Fernández, A. M. (2025). Back Home: A Computer Vision Solution to Seashell Identification for Ecological Restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5159-5168).

Wang, F., Li, H., Chao, W., Zhuo, Z., Ji, Y., Peng, C., & Sun, Y. (2025). E-convNeXt: A lightweight and efficient convNeXt variant with cross-stage partial connections. *arXiv preprint arXiv:2508.20955*.

Wu, L., & Wang, H. (2023). Global and pyramid convolutional neural network with hybrid attention mechanism for hyperspectral image classification. *Geocarto International, 38*(1), 2226112.

Xia, J., Yin, Y., & Li, X. (2025). An efficient medical image classification method based on a lightweight improved ConvNeXt-Tiny architecture. *arXiv preprint arXiv:2508.11532*.

Zhang, Z., Eli, E., Mamat, H., Aysa, A., & Ubul, K. (2023). EA-ConvNeXt: An approach to script identification in natural scenes based on edge flow and coordinate attention. *Electronics*, *12*(13), 2837.

Zheng, F., Lin, S., Zhou, W., & Huang, H. (2023). A lightweight dual-branch Swin Transformer for remote sensing scene classification. *Remote Sensing*, *15*(11), 2865.

Bibliographic information of this paper for citing:

Alkubaisi, Mohammed Hamid; Mishaal, Issa Mohammed; Sabri, Bassam Talib & Qasim, Kian Raheem (2026). ConvNeXt-Tiny for High-Accuracy Natural Scene Image Classification with Test-Time Augmentation. *Journal of Information Technology Management*, 18 (3), 228-242. https://doi.org/10.22059/jitm.2026.108385