Engineering Applications of Artificial Intelligence 162 (2025) 112654

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Driving efficiency in aerial scene classification: Insights from data augmentation, image processing, and multiscale Convolutional Neural Network models

Md Mahbub Hasan Rakib, Md Yearat Hossain, Rashedur M. Rahman *

Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

ARTICLE INFO

Keywords:
Lightweight deep learning
Image preprocessing
Gray-level co-occurrence matrix analysis
Multiscale feature extraction
Remote sensing

ABSTRACT

Aerial scene classification using satellite and drone imagery is vital for applications like environmental monitoring and urban planning, but the high computational demands of Convolutional Neural Networks (CNNs) limit their real-time use on resource-constrained platforms. Additionally, these models often lack global context awareness, relying mainly on small 3 × 3 kernels that capture fine details but miss broader spatial relationships. While large models can learn complex patterns, lightweight models struggle without adequate data and preprocessing. To address these challenges, we propose a lightweight CNN classifier optimized for fast, real-time aerial scene classification. Inspired by the Inception network, our architecture integrates multi-scale convolutional filters (1 × 1, 3 × 3, 5 × 5, and 7 × 7) to capture both local and global context. To reduce computational overhead, we incorporate Depthwise Separable Convolutions (DSC). To overcome the limitations of a compact model, we apply 5-fold semantic-preserving data augmentation and Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance feature visibility. While most state-of-the-art models require at least a million parameters, our lightweight CNN achieves an impressive 97.62 % accuracy on the UC (University of California) Merced Land Use Dataset using only 56,293 parameters. Furthermore, we use Gradient-Weighted Class Activation Mapping (Grad-CAM) to assess augmentation, CLAHE, and our architecture's influence on feature attention. This study demonstrates that effective preprocessing with CLAHE and extensive augmentation can narrow the gap between lightweight CNNs and complex models. These results support the use of efficient models for real-time, resource-constrained aerial scene classification, promoting sustainable and accessible Artificial Intelligence (AI) in remote sensing applications.

1. Introduction

Satellite and aerial imaging technologies have become more prevalent, providing large volumes of high-resolution imagery over vast geographic regions, thanks to the quick advancement of earth observation and remote sensing technologies. High spatial resolution imagery, offering sub-meter detail (Zhao et al., 2016), and very high-resolution remote sensing (RS) imagery with centimeter-level precision (Shawky et al., 2020), are now widely accessible. The global RS market is projected to grow rapidly, expanding from around $22–28 billion in the mid-2020s to over $150 billion by 2034 (Research). This surge reflects the increasing demand for aerial image analysis across environmental monitoring, urban planning, and agriculture applications. The increasing volume of RS imaging data has made aerial scene classification increasingly significant since it offers valuable insights into land use and land cover changes. This approach is crucial for applications such as urban planning and development (Wellmann et al., 2020), environmental monitoring (Padró et al., 2019; Yuan et al., 2020), agricultural management (Weiss et al., 2020; Huang et al., 2018), disaster management (Kucharczyk et al., 2021; Kemper et al., 2020; Bello et al., 2014), and land use mapping (Huang et al., 2020; Mohan-Rajan et al., 2020). Understanding land use and its temporal changes enables governments and organizations to monitor shifting patterns, detect deforestation, assess natural disaster impacts, and improve land resource management for sustainable development. The availability of advanced high-resolution imaging technologies calls for effective methods to classify aerial landscapes such as agricultural fields, forests, water bodies, and urban areas. Unlike natural images, RS images are

* Corresponding author.
E-mail address: rashedur.rahman@northsouth.edu (R.M. Rahman).

https://doi.org/10.1016/j.engappai.2025.112654
Received 23 December 2024; Received in revised form 17 August 2025; Accepted 4 October 2025

more complex, often containing diverse objects in a single frame (Zhang et al., 2016; Liu et al., 2021), with some ground features exhibiting high visual similarity (Tuia et al., 2011). As a result, current research emphasizes robust feature extraction techniques. The emergence of deep learning (DL) has significantly advanced computer vision, with Convolutional Neural Networks (CNNs) leading the way by automatically extracting and learning meaningful features such as edges, textures, and shapes directly from raw image data. Consequently, most recent studies rely on CNNs for effective aerial scene classification (Cheng et al., 2017; Luus et al., 2015; Zhang et al., 2015; Gui-Song et al., 2017; Liu et al., 2018).

Aerial scene classification has been widely studied using advanced feature extraction and attention-based techniques, reaching high accuracy. However, most models focus on performance rather than suitability for low-end edge devices (Xue et al., 2020; Wang et al., 2020; Wan et al., 2021; Zhang et al., 2019). This highlights the trade-off between accuracy and computational expense. Large CNNs may need hundreds of megabytes to several gigabytes of storage and can slow down inference on low-power processors (Liu et al., 2024; Tabani et al., 2021; Castanyer et al., 2021). They also consume significant battery energy, which is vital for mobile and wearable devices (Liu et al., 2024; Tabani et al., 2021; Pietrołaj et al., 2024). Furthermore, deployment can be difficult due to differences in hardware support, such as GPUs or NPUs. Often, real-time inference depends on offloading the process to external servers, which raises concerns about privacy, connectivity, and reliability (Tabani et al., 2021; Castanyer et al., 2021). To address the limitations of heavy models, some studies have proposed lightweight architectures for aerial scene classification (Shen et al., 2023; Bai et al., 2021; Yu et al., 2020). While large models can perform well even with small datasets, these lightweight models often struggle due to their shallow depth. Most of these lightweight networks utilize 3×3 kernels to reduce computation, but this limitation restricts their receptive field and weakens their ability to capture global context. This is a significant drawback for datasets like UC (Universuty of California) Merced Land Use Dataset, where several classes, such as golf courses, beaches, and agricultural fields, require understanding the entire scene rather than focusing on specific objects. Moreover, these works mainly focus on model design and pay little attention to data preprocessing. Effective data preprocessing increases variability in training data, enabling the model to better understand context and focus more accurately on important regions of an image.

To overcome these limitations, we propose a lightweight CNN classifier combined with effective preprocessing, including 5-fold semantic-preserving augmentation and Contrast Limited Adaptive Histogram Equalization (CLAHE) image enhancement techniques. Since aerial images contain subjects of varying sizes and both local and global contexts can be important in assessment, it is crucial to design CNNs that can capture and utilize these contexts for classification. To accomplish this, we have used a custom Inception module inspired by the Inception network (Szegedy et al., 2015). It uses 1×1, 3×3, 5×5, and 7×7 kernels to effectively capture both local and global contexts from images. Also, we have utilized depth-wise separable convolution (Chollet, 2017) operations in general to replace standard convolution filters to reduce the overall complexity of our network. Furthermore, we have utilized image augmentation and the CLAHE image processing technique. Numerous studies have been conducted to improve low-contrast images utilizing CLAHE in various domains, including medical, aerial, and underwater images (Malik et al., 2019; Vidhya et al., 2017; Harichandana et al., 2020; Salem et al., 2019; Santos et al., 2020; Garg et al., 2018). CLAHE improves image contrast, enhancing feature visibility and aiding model performance. The augmentation we used prevents semantic distortion and represents real-world aerial scenarios, where images can appear in any orientation. The image augmentation technique helps the model train on more data samples, while the CLAHE enhances image features, enabling lightweight models to learn features properly (Aboshosha et al., 2019). Even though CNNs perform remarkably well on tasks like object detection and image classification, they often function as "black boxes," providing little information about the decision-making processes behind these models. Selvaraju et al. (2017) introduced Gradient-Weighted Class Activation Mapping (Grad-CAM) to address this. It generates visual explanations by highlighting the regions of an input image that a model focuses on when making predictions, enabling researchers to better understand and diagnose model behavior. To investigate how image enhancement, data augmentation, and the proposed network affect the interpretability of lightweight CNN models, we also used Grad-CAM in this study. This approach reveals differences in feature extraction and decision-making, helping us understand the impact of preprocessing and network design on model attention.

The main contributions of this study are as follows:

1. We introduced a novel lightweight model architecture that utilizes Depthwise Separable Convolutions (DSC), hence decreasing parameters and computations, which results in accelerated inference, enabling our model to be suitable for deployment on mobile and edge devices.
2. We investigated the impact of data augmentation on a lightweight CNN architecture.
3. We explored the effect of the CLAHE image enhancement approach on a lightweight CNN architecture.
4. We employed Grad-CAM to analyze feature attention across models trained on datasets with and without CLAHE, demonstrating the impact of CLAHE on model interpretability, feature extraction, and performance.

The remainder of this paper is structured as follows: Section 2 reviews related works and highlights the research gap. Section 3 details the proposed methodology. Section 4 describes the experimental setup and presents the performance evaluation. Section 5 discusses the results, and Section 6 concludes the study with directions for future research.

2. Related works

Researchers are exploring various methods to improve aerial scene classification accuracy, with a key focus on enhancing feature extraction, as features play a crucial role in classification performance.

Shen et al. (2023) proposed a modified lightweight architecture based on GhostNet, designed for real-time scene classification in embedded and resource-limited environments. On the University of California Merced Land Use Dataset (UC Merced) dataset and similar benchmarks, the modified GhostNet greatly reduced computational complexity from 7.85 million to 2.58 million FLOPS and memory usage from 16.4 MB to 5.7 MB while achieving a higher classification accuracy of 96.19 %, outperforming its original version as well as MobileNetV3-Small. Although their proposed method is lightweight, it lacks multiscale feature extraction capabilities, which is a notable limitation when working with datasets like UC Merced that require a strong global contextual understanding. Furthermore, while their data augmentation strategy includes rotation, which is generally beneficial, the addition of brightness and contrast adjustments may alter semantic information, potentially introducing noise and negatively affecting classification performance.

Xue et al. (2020) proposed a classification technique that utilizes multi-structure deep feature fusion (MSDFF). First, they used random-scale cropping for data augmentation. After that, they employed three distinct CNNs, GoogLeNet, CaffeNet, and VGG-VD16, to extract deep features from images, and then they used these features to fuse using a deep feature fusion network for classification. Their proposed method, trained on 80 % of the UC Merced dataset, achieved 99.76 % accuracy. However, their proposed architecture uses around 134 million parameters, which makes the model computationally expensive.

Table 1
Comparison of existing methods concerning image enhancement, semantic-preserving augmentation, and edge device deployability.

Study	Backbone	Preprocessing		Edge Device Deployable	Limitation
		Incorporation of Image Enhancement	Semantic-preserving Data Augmentation		
Sheen et al. (Shen et al., 2023)	GhostNet	No	No (Brightness, Contrast)	Yes	Lacks multiscale feature extraction
Xue et al. (Xue et al., 2020)	GoogLeNet, CaffeNet, VGG-VD16 (MSDFF)	No	No (Random-scale cropping)	No	Extremely high parameter count (~134M)
Wang et al. (Wang et al., 2020)	CNN + Saliency Fusion	No	No	Yes	No augmentation; saliency fusion only
Bai et al. (Bai et al., 2021)	ESPA-MSDWNet (MobileNet V2)	No	No	Yes	Relatively high parameter usage (2.39M)
Wan et al. (Wan et al., 2021)	ResNeXt	No	Yes (Horizontal flip, random rotation)	No	High parameter count (~25M)
Zhang et al. (Zhang et al., 2019)	SE-Net	No	No	No	High parameter count (~19.72)
Yu et al. (Yu et al., 2020)	MobileNet V2 + Dilated Conv	No	Yes (rotation, flip)	Yes	No CLAHE; could improve with better image enhancement
Chen et al. (Chen et al., 2025)	MLCMFNet (Swin Transformer)	No	Yes (conditional GAN [53], and self-attention GAN	No	High parameter count (~35M)
Shi et al. (Shi et al., 2025)	RepFACNN (CNN + Transformer)	No	Yes (Mixup)	Yes	Computationally expensive (~6.4M)
Proposed Work	Custom Lightweight CNN	Yes (CLAHE)	Yes	Yes	Aims to match heavy models' accuracy with a lightweight architecture

To get more informative features, Wang et al. (2020) used a saliency detection model that mimics the human selective visual attention mechanism to process the images. They used a feature fusion approach, combining two feature sets: one generated from RGB images and the other from processed images using saliency detection. They attained an overall accuracy of 87.15 ± 0.69 % on RGB images without the fusion approach, and 98.04 ± 0.89 % using the proposed fusion method. They also highlighted the model's lightweight nature while embracing the benefits of feature fusion and multiscale techniques. Yet, the method doesn't utilize any augmentation strategies, which would help make the model more generalized and robust.

In order to achieve high accuracy and inexpensive model parameters, Bai et al. (2021) presented ESPA-MSDWNet, a lightweight multiscale depthwise network with effective spatial pyramid attention (ESPA). It increases receptive fields and uses depthwise convolution to obtain granular multiscale features using MobileNet V2 as the backbone. However, their proposed model contains around 2.39 million parameters, which suffer from inefficiency.

Using ResNeXt as the backbone, the authors of the study (Wan et al., 2021) proposed LmNet, which combines multiscale feature fusion and lightweight channel attention to help learn important channel features quickly. Additionally, they proposed a multiscale feature fusion framework that integrates both shallow edge and deep semantic information to enhance feature representation and contribute to classification accuracy. Nevertheless, their methodology results in a model that uses around 25 million parameters. Due to this high quantity of parameters, the model becomes computationally expensive.

Zhang et al. (2019) introduced a lightweight and effective CNN utilizing MobileNet V2, integrating dilated convolution and channel attention to enhance feature extraction. To improve efficiency, they implemented a multidilation pooling module to capture multiscale characteristics, hence ensuring excellent accuracy. Although they used image augmentation, the application of the CLAHE method could help the model to learn the intricate details in the images, making it more robust.

Yu et al. (2020) employed MobileNetv2 as the backbone model to extract deep image features, which are subsequently processed by two separate convolutional layers. The modified features go through a Hadamard product operation to produce enhanced bilinear features, which are subsequently pooled, normalized, and utilized for classification. This model also suffers from a high number of model parameters, which is around 7.76 million.

Chen et al. (2025) introduced MLCMFNet, a mutual learning method that combines multiple types of features. The network consists of three main components: a Multi-Feature Fusion Module (MFFM), which adds additional fused features; a Swin Transformer Module (STM), which captures global features; and a Multi-Attention Fusion Module (MAFM), which extracts more representative features. Mutual learning helps two different networks learn from each other, combining their strengths. This improves how local and global features work together. Their proposed model scored 99.91 ± 0.1 % accuracy on the UC Merced dataset. However, the model contains 35 million parameters; therefore, it may fail to operate properly in resource-constrained environments.

Shi et al. (2025) presented RepFACNN, a network that combines CNNs and transformers while reducing computation by using a reparametrized transformer (RepFormer). It extracts multilevel and multi-scale spatial features, then fuses them for better classification. This method scored 99.33 ± 0.26 % accuracy on the UC Merced dataset. Nevertheless, this model is computationally quite expensive as it utilizes almost 6.4 million parameters.

In contrast to prior studies, our approach focuses on designing a lightweight classification model with significantly fewer parameters and reduced computational requirements, while incorporating superior multiscale feature extraction to meet the demands of complex datasets. Additionally, we introduce a robust preprocessing pipeline that not only enhances image quality but also increases data diversity, effectively supporting the lightweight model in narrowing the performance gap with larger architectures. The research gap compared to existing methods is summarized in Table 1.

3. Methodology

This section discusses the datasets and techniques used for aerial scene classification. We split the UC Merced Land Use Dataset into training and testing sets and then applied data augmentation techniques to enhance generalization. Next, we employed the CLAHE method to improve the image's contrast, and we implemented a custom CNN architecture that integrates inception modules, GAP, and DSC to facilitate efficient feature extraction and classification. Subsequently, we employed Grad-CAM for model interpretability. Finally, we evaluated Gray-Level Co-occurrence Matrix (GLCM) features to see how the original and CLAHE-processed images' textures differed. Finally, we assessed the model's performance using metrics like overall accuracy and a confusion matrix to compare the effectiveness of the original and

Table 2
Detailed information about the UC Merced Land Use Dataset.

Dataset	Number of classes	Number of images per class	Total number of images	Image size	Spatial resolution (m)
UC-Merced	21	100	2100	256 × 256	0.3

Fig. 1. Sample images of each class in UC-Merced dataset: (1) agricultural, (2) airplane, (3) baseball diamond, (4) beach, (5) buildings, (6) chaparral, (7) dense residential, (8) forest, (9) freeway, (10) golf course, (11) harbor, (12) intersection, (13) medium residential, (14) mobile home park, (15) overpass, (16) parking lot, (17) river, (18) runway, (19) sparse residential. (20) storage tanks, (21) tennis courts.

processed datasets.

3.1. Datasets

Deep learning relies heavily on datasets since they provide the basis for model training. Repeated exposure to data during training allows deep learning algorithms to discover patterns, features, and relationships. A model can only learn well with adequate and pertinent data, which produces subpar performance and generalization. A well-curated dataset allows the model to make accurate predictions by offering a variety of instances that reflect the real-world variables the model is likely to face. The quality and quantity of data substantially impact the model's success. As a result, for this investigation, we employed the UC Merced Land Use Dataset (Yang and Newsam). Detailed information on the UC-Merced dataset is given in Table 2. The UC Merced Land Use Dataset, released by the University of California, Merced, is a widely used benchmark in remote sensing and land use classification. This dataset comprises 2100 high-resolution RGB aerial images, each of size 256x256 pixels. The dataset includes 21 land use classes, such as agricultural, airplane, beach, forest, and residential areas. The images,

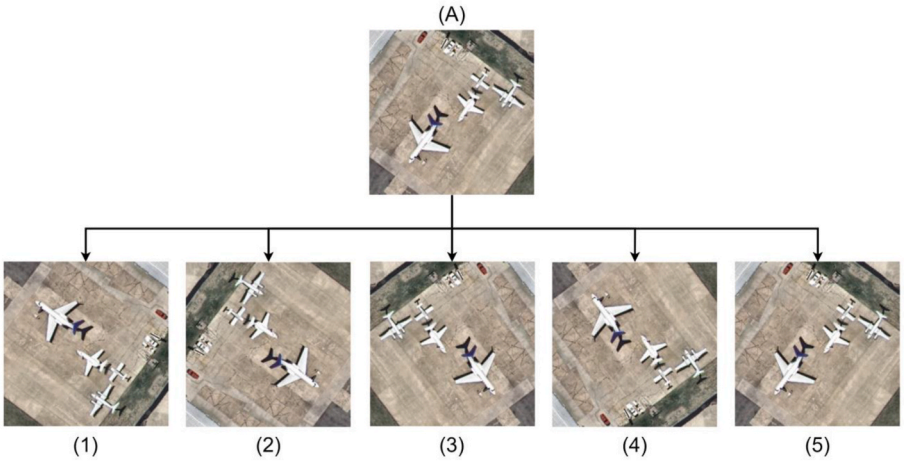

Fig. 2. Augmented variations of original images using 5-factor offline augmentation. Where (A) is the original image from the dataset. After applying 5-factor augmentation, we get augmented images. These are: (1) Rotation right by 90°, (2) Rotation left by 90°, (3) Flip horizontal, (4) Flip vertical, and (5) Identity (no augmentation).

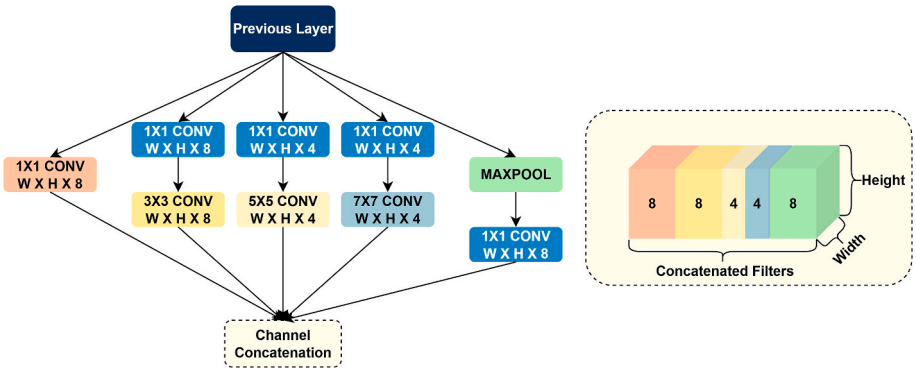

Fig. 3. A custom inception module with the addition of an extra 7×7 kernel to the original inception module.

sourced from the USGS National Map Urban Area Imagery, have a spatial resolution of 0.3 m per pixel and represent various natural and human-made environments. Sample images from each category in this dataset are shown in Fig. 1.

3.2. Data augmentation

Augmentation is vital in deep learning because it increases the training data's diversity and amount, which helps minimize overfitting and improves the model's generalization capacity. Particularly in the case of small datasets like the UC-Merced land use dataset, with only 100 images per class, augmentation introduces variations such as changes in orientation and reflections, allowing the model to handle the complexities of real-world data better. By transforming images without changing their semantic meaning, the model can learn to identify patterns and features more effectively, even when meeting previously unknown variances in test data.

In this work, we used offline augmentation, where augmentations are applied before training, rather than online augmentation, which applies transformations during training. Offline augmentation ensures that the augmented data is fixed and predefined, providing a consistent dataset throughout multiple experiments, which is notably useful for reproducibility in tasks such as model comparison and ablation studies. We applied a 5-factor augmentation strategy illustrated in Fig. 2, where each image $x \in D_{train}$ is subjected to the following transformation: (1) rotated right by 90°, using $Rotate_{90}(x)$ operator that rotates the image by 90° clockwise, (2) rotated left by 90°, using $Rotate_{-90}(x)$ operator that rotates the image by 90° counterclockwise, (3) flipped horizontally, using $Flip_{horizontal}(x)$ operator that flips the image horizontally, (4) flipped vertically using the $Flip_{vertical}(x)$ operator that flips the image vertically, and (5) Identity (no changes) using $Identity(x)$ operator that leaves the image unchanged. Thus, each image $x_i \in D_{train}$ generated an augmented set $A(x_i) = \{x_{identity}, x_{rotate-right}, x_{rotate-left}, x_{flip-horizontal}, x_{flip-vertical}\}$. By using these augmentations, we expanded the training set fivefold, yielding a more diverse and robust dataset.

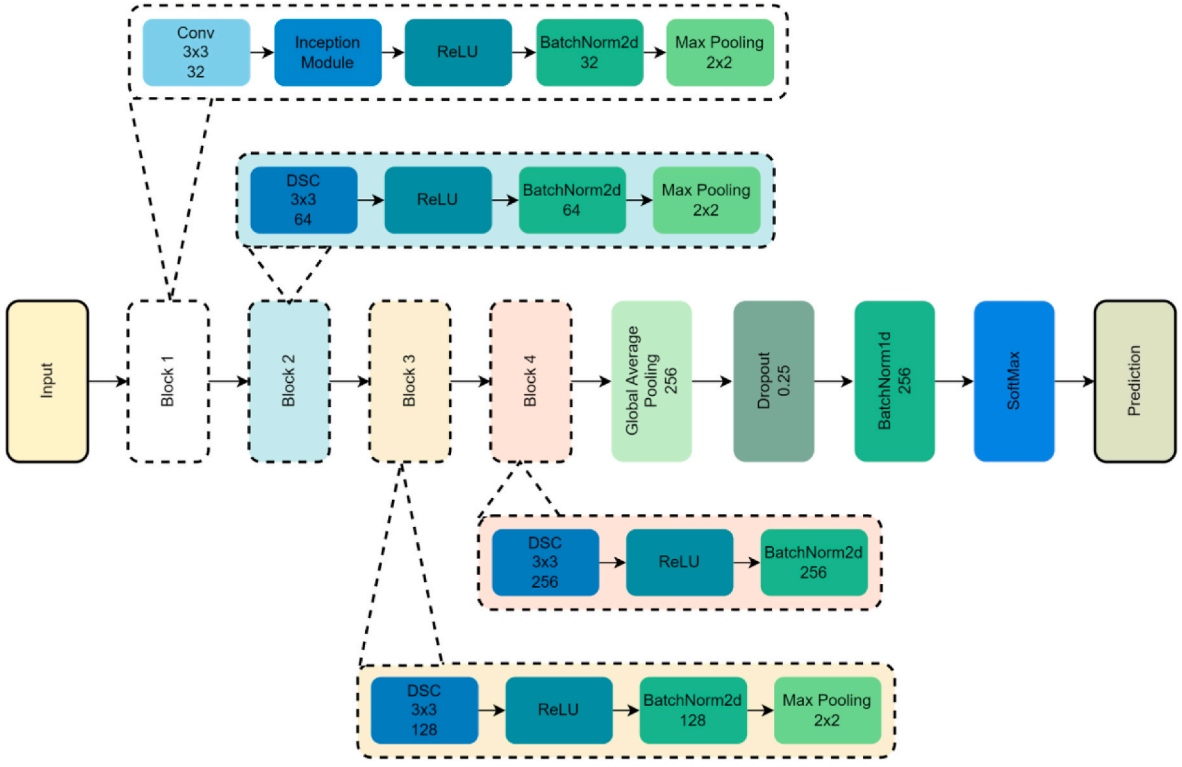

Fig. 4. Overview of the proposed architecture for aerial scene classification. The architecture integrates Inception modules and DSC to enhance feature extraction while maintaining computational efficiency.

Table 3
A detailed summary of our proposed design, specifying each layer type, the associated output dimensions, and the number of parameters related to each layer.

Layer (type)	Convolutional Layer Parameters			Output Shape			Parameters
	kernel_size	Padding	Stride	Channel	Height	Width	
Conv2d	*3*	*1*	*1*	*32*	*256*	*256*	*896*
Conv2d (Inception Module)	1	0	1	8	256	256	264
Conv2d (Inception Module)	1	0	1	8	256	256	264
Conv2d (Inception Module)	3	1	1	8	256	256	584
Conv2d (Inception Module)	1	0	1	4	256	256	132
Conv2d (Inception Module)	5	2	1	4	256	256	404
Conv2d (Inception Module)	1	0	1	4	256	256	132
Conv2d (Inception Module)	7	3	1	4	256	256	768
MaxPool2d (Inception Module)	3	1	1	32	256	256	0
Conv2d (Inception Module)	1	0	1	8	256	256	264
Inception Module	N/A	N/A	N/A	32	256	256	0
ReLU	N/A	N/A	N/A	32	256	256	0
BatchNorm2d	N/A	N/A	N/A	32	256	256	64
MaxPool2d	2	0	2	32	128	128	0
Conv2d (Depthwise)	3	1	1	32	128	128	320
Conv2d (Pointwise)	1	0	1	64	128	128	2112
ReLU	N/A	N/A	N/A	64	128	128	0
BatchNorm2d	N/A	N/A	N/A	64	128	128	128
MaxPool2d	3	1	2	64	64	64	0
Conv2d (Depthwise)	3	1	1	64	64	64	640
Conv2d (Pointwise)	1	0	1	128	64	64	8320
ReLU	N/A	N/A	N/A	128	64	64	0
BatchNorm2d	N/A	N/A	N/A	128	64	64	256
MaxPool2d	2	0	2	128	32	32	0
Conv2d (Depthwise)	3	1	1	128	32	32	1280
Conv2d (Pointwise)	1	0	1	256	32	32	33024
ReLU	N/A	N/A	N/A	256	32	32	0
BatchNorm2d	N/A	N/A	N/A	256	32	32	512
AdaptiveAvgPool2d	N/A	N/A	N/A	256	1	1	0
Dropout	N/A	N/A	N/A	256	N/A	N/A	0
BatchNorm1d	N/A	N/A	N/A	256	N/A	N/A	512
Softmax	N/A	N/A	N/A	21	N/A	N/A	5397

Total Parameters: 56293.
Trainable Parameters: 56293.
Non-Trainable Parameters: 0.

3.3. CLAHE

CLAHE is a cutting-edge image processing technique that improves image contrast, especially in cases where global contrast adjustments might result in over-saturation of regions of relatively uniform intensity. Unlike standard Histogram Equalization (HE), which adjusts contrast uniformly across the entire image, CLAHE works on small, localized sections known as tiles. To avoid obvious artifacts, the borders between the tiles are smoothly blended after localized equalization. One important characteristic of CLAHE is its ability to limit contrast amplification by establishing a clip limit, which keeps noise from being excessively amplified in areas of the image that are homogeneous. Because of this, CLAHE works particularly well for bringing out the subtle details in images with varying contrast, as those seen in remote sensing, medical imaging, and other applications where retaining subtle details is crucial while avoiding excessive noise amplification. Hence, we applied CLAHE with a clip limit of 2 and a tile grid size of 8×8 to generate an enhanced version of the dataset. Given an image $I(x,y)$, the image is divided into $N \times M$ tiles, and for each tile, a histogram is computed. The contrast is limited by clipping the histogram at a specified threshold T, where any bin $H(i)$ in the histogram exceeding T is clipped, and the excess is redistributed across the remaining bins. The clipped histogram is then used to map the pixel values through a cumulative distribution function that is expressed as follows:

$$(CDF), C(i) = \sum_{j=0}^{i} H(j) \tag{1}$$

Equation (1) is used to transform the pixel intensities. The final output for each pixel $I'(x,y)$ is obtained by bilinear interpolation between neighboring tiles to avoid boundary artifacts. Thus, CLAHE enhances local contrast while suppressing noise amplification.

3.4. Proposed architecture

In this section, we provide an in-depth description of the architecture we are proposing. As shown in Fig. 4, the initial block of the model incorporates a convolutional block, which takes 3 input channels and outputs a feature map of 32 channels, followed by an Inception module, a core element of the Inception architecture (Szegedy et al., 2015), with 32 channels. The Inception module enables multiscale feature extraction by applying multiple convolutional filters (1×1, 3×3, 5×5) and a max-pooling operation in parallel. A 1×1 convolution is used before larger filters to reduce dimensionality and computational cost. While 3×3 filters capture fine details, 5×5 filters extract more global features. The results of these operations are concatenated along the depth (channel) dimension, yielding a rich feature map made up of information retrieved at multiple scales. This enables the network to capture features at various sizes without committing to a particular convolution size, resulting in improved performance while keeping computational costs under control. By concatenating features from distinct branches, the network can adapt to various patterns, allowing it to better generalize to diverse datasets, such as remote sensing images with both small and large structures. Drawing inspiration from the original Inception module, we added a 7x7 kernel to the original module for our objective of aerial scene classification.

The architecture of the custom inception module is shown in Fig. 3. The reason behind incorporating a 7x7 kernel stems from the nature of the UC-Merced dataset, which includes land image categories such as golf courses, sparse residential areas, and dense residential areas, as

illustrated in Fig. 1. In such scenes, patterns or objects are often dispersed across larger regions. Larger kernels are beneficial in these cases, as they can capture distant spatial relationships. Specifically, a 7 × 7 kernel, due to its larger receptive field, can extract broader contextual features and capture more large-scale or global patterns within an image. For aerial scene classification, this can enhance the model's ability to identify entire regions, such as forests, agricultural fields, or large residential zones.

In the second, third, and fourth blocks, we employed DSC instead of normal convolution, aiming to significantly reduce the computational load of the network without sacrificing performance. This approach reduces computation in CNNs by decomposing a standard convolution into two operations: depthwise and pointwise convolutions. The depthwise step applies one filter per input channel, reducing operations from $K \times K \times C_{in} \times C_{out} \times H \times W$ to $K \times K \times C_{in} \times H \times W$. Then, the pointwise ($1 \times 1$) convolution combines these outputs with $C_{in} \times C_{out} \times H \times W$ operations. This significantly lowers the computational cost, ideal for real-time and resource-constrained applications.

The model employs batch normalization to stabilize training and utilizes ReLU activation to introduce nonlinearity. The network concludes with a Global Average Pooling (GAP) (Lin, 2013) layer followed by a dropout layer to mitigate overfitting. GAP aggregates spatial data throughout the entire feature map, reducing each channel into one value. It calculates the average value of each feature map over all spatial dimensions, effectively reducing each feature map to a single value. Mathematically, for a given feature map F of size $H \times W$, the output y after applying GAP can be expressed as:

$$y = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F(i,j) \tag{2}$$

Where $F(i,j)$ denotes the value at the i-th row and j-th column of the feature map. The resulting output y is a vector of size C, where C is the number of feature maps. This approach not only preserves the global context of the image but also significantly decreases the total number of trainable parameters to 56,293, as indicated in Table 3. Furthermore, GAP is particularly good for reducing overfitting in deep networks since it introduces fewer parameters than fully connected layers, making it a great choice for transferring information across different input sizes and improving the model's performance. The final output layer employs softmax activation to generate probabilities for each class, making it ideal for multi-class classification tasks.

Our proposed model architecture balances complexity and performance by incorporating Inception modules, DSC, and a GAP layer. The Inception module utilizes parallel convolutional filters of varying sizes to capture multi-scale features, enabling the model to generate broad feature representations while reducing the likelihood of overfitting. DSC decreases parameter counts and computational complexity, thereby improving training and inference speed. The GAP layer reduces spatial dimensions by aggregating feature maps into single vectors, reducing the risk of overfitting while preserving critical global context. This combination produces a lightweight and powerful model demonstrating efficient learning and high accuracy in aerial scene classification tasks.

3.5. Grad-CAM

Grad-CAM is a method for visualizing and explaining the decision-making process of a CNN by highlighting the regions of an image that are significant for class prediction. The technique employs the gradients of the output from the feature maps of the last convolutional layer to produce a heatmap highlighting the significant regions influencing class activation. The procedure starts with a forward pass to derive the class score $S(c)$ for the class c, subsequently calculating the gradients of this score regarding the feature maps A^k of the final convolutional layer, denoted as $\frac{\delta S(c)}{\delta A^k}$. After that, GAP is utilized to calculate the weights for each feature map using Equation (3).

$$a_k = \frac{1}{Z} \sum_i \sum_j \frac{\delta S(c)}{\delta A^k(i,j)} \tag{3}$$

Where Z represents the number of pixels in the feature map. The Grad-CAM heatmap L is produced by integrating the feature maps with their respective weights, expressed as $L = ReLU\left(\sum_k a_k A^k\right)$, whereas ReLU guarantees that only positive values are taken into account. This heatmap is then upsampled to the original image size, which improves CNN interpretability by enabling the presentation of the crucial areas that affect the model's classification. As a powerful tool for model interpretation, Grad-CAM offers valuable insights into CNN decision-making by graphically representing the regions of the input image that are most significant for a particular classification.

3.6. Gray-level Co-occurrence matrix (GLCM)

GLCM, introduced by Haralick et al. (1973), is a widespread statistical method in image processing for analyzing image texture. This technique assesses the spatial correlations among pixel intensities, providing insights about the image's structural patterns. The GLCM counts the frequency of co-occurrence of pixel value pairs with predefined intensities, known as gray levels, within a designated spatial arrangement in a chosen image region. It measures the frequency of 2 gray levels i and j occurring at a particular spatial distance d and orientation θ. The GLCM, denoted as, $P(i,j)$ is formed by determining the occurrences of pixel pairs (i,j), where i signifies the gray level of a reference pixel and j indicates the gray level of its adjacent pixel, situated at a specified distance d and angle θ. Mathematically, $P(i,j)$ is defined as:

$$P(i,j|d,\theta) = \sum_{x=1}^{M} \sum_{y=1}^{N} \begin{Bmatrix} 1, if\ I(x,y) = i\ and\ I(x+d_x y+d_y) = j \\ 0, otherwise \end{Bmatrix} \tag{4}$$

Where $I(x,y)$ is the intensity at the pixel (x,y), and (d_x, d_y) is the offset corresponding to the specified d and θ. The result is a matrix where each element $P(i,j)$ represents the number of times the pair (i,j) occurs in the image with the given spatial relationship. From GLCM, several texture features can be derived, such as contrast, correlation, energy, and homogeneity. Contrast measures the intensity difference between a pixel and its neighbor. It is calculated using Equation (5).

$$Contrast = \sum_{i,j} P(i,j)(i-j)^2 \tag{5}$$

Correlation measures how correlated a pixel is to its neighbor, which is calculated using Equation (6).

$$Correlation = \sum_{i,j} \frac{(i-\mu_i)\left(j-\mu_j\right)P(i,j)}{\sigma_i \sigma_j} \tag{6}$$

Where μ and σ are the mean and standard deviation of the gray levels.

Energy represents the uniformity of the texture. Equation (7) is used to calculate energy.

$$Energy = \sum_{i,j} P(i,j)^2 \tag{7}$$

Homogeneity measures the closeness of the distribution of elements in the GLCM to the diagonal. It is calculated using Equation (8).

$$Homogeneity = \frac{P(i,j)}{1+|i-j|} \tag{8}$$

GLCM is an effective texture analysis tool that reveals spatial patterns of pixel intensities in an image.

Fig. 5. Sample preview of the original and CLAHE-processed images from three different land use classes in the UC Merced Land Use Dataset. The top row includes the original images for the following classes: airplane, baseball diamond, and storage tanks. The bottom row shows the CLAHE-processed images for the same classes.

3.7. Metrics

Model complexity and classification performance were the two evaluation criteria for our proposed method. Classification performance comprises overall accuracy and the confusion matrix, whereas model complexity involves FLOPS, inference times, and size.

3.7.1. Overall accuracy

Overall Accuracy (OA) is the ratio of correct predictions to the total number of predictions made. It gives an overall measure of how well the model performs across all classes. Mathematically, OA can be expressed as:

$$OA = \frac{1}{T} \sum_{i=1}^{n} \sum_{j=1}^{k} P_{ij} \tag{9}$$

Where P_{ij} represents the count of accurately predicted instances, T denotes the total examples in the test set, n signifies the number of instances per class, and k indicates the total number of categories or classes.

3.7.2. Confusion matrix

A confusion matrix is used to evaluate the effectiveness of a classification model. This provides a better understanding of the performance by displaying the number of predictions that were correct and those that were incorrect for each class. The matrix is useful for determining the different kinds of errors the model is making, as well as the percentage of incorrect classifications that occur across the various classes. If a multi-class instance involves n classes, the confusion matrix will be a square of size $n \times n$, where each entry P_{ij} represents the number of instances where the true label is class i and the predicted label is class j.

3.7.3. Floating point operations per second (FLOPS)

The amount of computational complexity that a model comes with can be measured using FLOPS. It indicates the number of floating-point operations that are carried out by the model during the inference process. The model is generally considered to be more efficient when it has lower FLOPS.

3.7.4. Inference time

Inference Time measures the time it takes for the model to make predictions on new data. This can be evaluated by running the model on a set of test images and recording the time taken for the forward pass.

3.7.5. Size

Model Size refers to the model's saved weights or parameters. This can be measured by checking the storage size of the model file. This is important for deployment in resource-constrained environments.

4. Experiments and results

In this section, we discuss the experimental setup and results obtained from the experiments that were conducted. The experimental setup outlined the training environment and parameters for the custom CNN model. We described the training process, dataset splits, and evaluation strategy in the experimental details. We explored the impact

Table 4
UC Merced Land Use Dataset distribution with and without augmentation.

Dataset	Split Ratio	Number of images per class		Number of images per set		Total images in the dataset
		Train set	Test set	Train set	Test set	
Without augmentation	80:20	80	20	1680	420	2100
With 5-fold augmentation	80:20	400	20	8400	420	8820

of data augmentation and CLAHE enhancement, demonstrating their influence on classification performance. A comparison with other work in remote sensing classification was provided, showing the advantages of our proposed approach. Grad-CAM visualizations were presented to interpret the model's decision-making process by highlighting important regions in both original and CLAHE-enhanced images.

4.1. Experimental setup

All experiments were conducted using the PyTorch framework (version 1.13.1) with Python 3.10.6. Training was performed on a system equipped with an Intel Core i7-13700K CPU, 32 GB DDR5 RAM, and an NVIDIA RTX 4090 GPU running Windows 11. The CUDA and cuDNN versions used were 11.7 and 8.4, respectively. To ensure reproducibility and deterministic behavior, we set the random seed to 42 across all relevant libraries: random.seed(42), np.random.seed(42), torch.manual_seed(42), and torch.cuda.manual_seed_all(42). Additionally, torch.backends.cudnn.deterministic was set to True. The models were trained for 500 epochs using a batch size of 64. The optimizer used was Adam with an initial learning rate of 0.01, while all other optimizer parameters were kept at their default values. A cosine annealing learning rate scheduler (CosineAnnealingLR) was used, with T_max set to 500 and eta_min set to 1×10^{-8}, while other parameters remained at their default values. The loss function used was the standard CrossEntropyLoss.

4.2. Experimental details

This study uses the UC Merced Land Use Dataset, widely used in remote sensing scene classification. As shown in Table 2, the dataset consists of 2100 images across 21 land-use classes, with 100 images per class. These classes represent a diverse range of environments, including residential areas, airports, beaches, etc. To investigate the impact of contrast enhancement on model performance, this study applied CLAHE preprocessing to the dataset. The processed images were then used alongside the original to evaluate whether CLAHE improves classification performance in lightweight CNNs. Fig. 5 compares original and CLAHE-processed images, highlighting improved local contrast and more uniform color distribution in the latter. These visual differences suggest that CLAHE may enhance feature extraction and help the model better distinguish between land-use categories.

To prepare the data for model training and evaluation, we split it into two subsets at an 80:20 ratio. Specifically, 80 % of the data was allocated to the training set, while the remaining 20 % was reserved for the test set to assess performance. This split was intended to achieve a balance between sufficient training data for model learning and ample test data for reliable generalization evaluation. After splitting the data into training and test sets, we applied data augmentation to both the original and the CLAHE-processed datasets. However, the augmentation was only conducted on the training set, leaving the test set unchanged to ensure a fair evaluation. Table 4 shows the image counts for each class, highlighting the number of images before and after data augmentation. This comparison shows the scale of augmentation applied to the training set, emphasizing the substantial increase in training data through augmentation, which was used to improve model performance and robustness across both the original and CLAHE-processed datasets.

Table 5
Comparison of three distinct models based on key performance and efficiency metrics. The metrics include FLOPS, Inference Time, and Model Size. The original dataset without any CLAHE processing is used to train Model A, the augmented dataset without any CLAHE processing is used to train Model B, and the dataset that has undergone CLAHE processing and augmentation is used to train Model C.

Model	GFLOPS	Inference Time (seconds per image)		Size (KB)
		CPU (i7-13700K)	GPU (RTX 4090)	
Model A	0.3519	0.033546 ± 0.00003	0.003201 ± 0.00002	244
Model B	0.3519	0.033564 ± 0.00002	0.003212 ± 0.00002	244
Model C	0.3519	0.033867 ± 0.00003	0.003492 ± 0.00003	244

To evaluate the impact of data augmentation and CLAHE processing on the performance of our proposed lightweight CNN, we trained the same architecture under three different conditions. Each model was trained for 500 epochs using a cosine annealing learning rate scheduler to support smooth convergence. By keeping the architecture and training settings consistent, performance differences could be attributed solely to the effects of augmentation and CLAHE. A summary of the proposed architecture is provided in Table 3.

The first model, serving as the baseline, was trained on the original dataset without any augmentation or CLAHE. The second model incorporated 5-fold offline augmentation to improve generalization, as illustrated in Fig. 2, but excluded CLAHE to isolate the effect of augmentation. The third model was trained using a dataset processed with both 5-fold augmentation and CLAHE. All three models share the same architecture and thus have identical FLOPS and model sizes. However, inference time may vary slightly due to hardware conditions. To obtain consistent timing results, we measured inference time over 10 runs on a test platform equipped with an NVIDIA RTX 4090 GPU and a 13th Gen Intel Core i7-13700K CPU. Averaging the results helped minimize fluctuations and provide a reliable assessment. The mean inference times are presented in Table 5. We measured the inference time of the proposed model on the UC Merced dataset using a batch size of 1 on an NVIDIA RTX 4090 GPU. After 2 warmup runs, inference was timed over 10 runs. The average inference time was mean_time $\pm$ std_time seconds per image.

4.3. Impact of data augmentation and CLAHE enhancement

In order to thoroughly evaluate the quality enhancements brought about by CLAHE, we employed two primary methods. First, we conducted histogram analysis across the red, green, and blue channels. This enabled us to see how applying CLAHE caused the pixel intensities to redistribute. As can be seen in Fig. 6, the histograms of the original images ((a) and (c)) exhibit narrow and concentrated peaks, particularly in the blue and green channels, indicating a restricted dynamic range and limited color variance. This pattern suggests that large homogeneous regions dominate the images, such as the concrete surface surrounding the airplane or the grassy field of the baseball diamond. Consequently, large portions of the images possess nearly uniform intensities, leading to reduced contrast and weak differentiation between neighboring regions. This, in turn, implies that certain details, such as shadows, textures, or transitions between objects like grass and the baseball diamond, are not well-defined. As a result, classification algorithms might struggle to detect features accurately in these regions. After applying CLAHE, the histograms of images ((b) and (d)) exhibit a more uniform distribution of pixel intensities across all three channels. This indicates that the contrast has been enhanced by distributing intensities evenly and making subtle features, such as varied shades of green in the grass or the outer edges of objects, more apparent. The equalization in all channels implies that CLAHE has distributed pixel

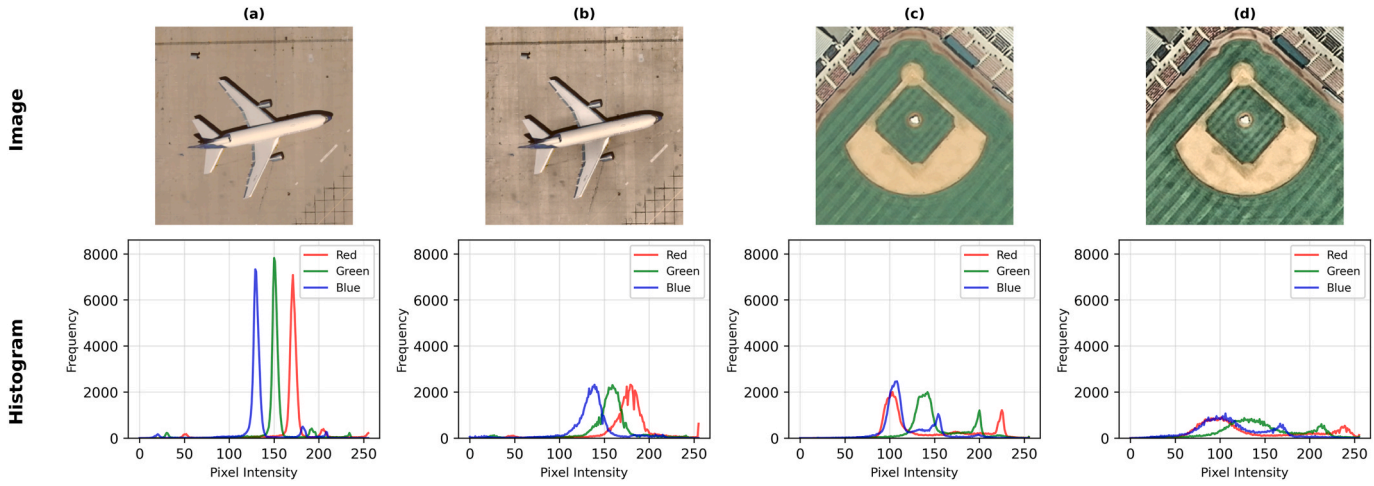

Fig. 6. Comparison of original and CLAHE-processed images with corresponding histograms. Row 1 shows image samples, where columns (a) and (c) are original images, and columns (b) and (d) are their CLAHE-processed counterparts. Row 2 shows the histograms, with (a) and (c) illustrating the pixel intensity distributions across red, green, and blue channels of the original images, and (b) and (d) showing the enhanced contrast and redistributed pixel intensities after processing with CLAHE. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 6
Comparison of GLCM Feature Analysis between CLAHE-Processed and original Datasets across 21 Classes of the UC-Merced Land Use Dataset.

Class	Contrast		Correlation		Energy		Homogeneity	
	Original Dataset	CLAHE Processed Dataset	Original Dataset	CLAHE Processed Dataset	Original Dataset	CLAHE Processed Dataset	Original Dataset	CLAHE Processed Dataset
agricultural	368.95	1514.26	0.64	0.63	0.02	0.01	0.11	0.05
airplane	280.13	523.56	0.91	0.88	0.03	0.01	0.22	0.11
baseball diamond	108.74	338.67	0.94	0.90	0.03	0.02	0.24	0.11
beach	84.62	323.27	0.96	0.91	0.06	0.03	0.38	0.19
buildings	509.03	811.62	0.94	0.90	0.02	0.01	0.19	0.10
chaparral	355.88	1353.88	0.81	0.80	0.02	0.01	0.11	0.06
dense residential	463.60	921.61	0.90	0.88	0.02	0.01	0.15	0.09
forest	265.80	1232.49	0.78	0.76	0.02	0.01	0.11	0.06
freeway	248.45	677.18	0.92	0.88	0.02	0.01	0.16	0.08
golf course	127.09	512.71	0.91	0.83	0.03	0.01	0.20	0.09
harbor	732.45	981.17	0.90	0.88	0.05	0.02	0.24	0.13
intersection	415.32	869.38	0.89	0.86	0.02	0.01	0.17	0.09
medium residential	475.41	1059.80	0.90	0.86	0.02	0.01	0.15	0.09
mobile home park	844.75	1311.57	0.88	0.86	0.02	0.01	0.16	0.10
overpass	436.13	937.05	0.89	0.84	0.02	0.01	0.16	0.09
parking lot	890.21	1360.43	0.83	0.81	0.02	0.01	0.13	0.08
river	289.93	1082.17	0.87	0.79	0.03	0.01	0.16	0.08
runway	254.52	546.70	0.90	0.85	0.04	0.02	0.23	0.12
sparse residential	287.99	877.43	0.90	0.84	0.02	0.01	0.15	0.08
storage tanks	526.72	898.49	0.89	0.85	0.03	0.02	0.23	0.13
tennis court	315.80	799.76	0.91	0.86	0.02	0.01	0.18	0.09

Table 7
Mean and Standard Deviation of GLCM Features for CLAHE-Processed and original Datasets across 21 Classes of the UC-Merced Land Use Dataset.

Dataset	Contrast		Correlation		Energy		Homogeneity	
	Original Dataset	CLAHE Processed Dataset	Original Dataset	CLAHE Processed Dataset	Original Dataset	CLAHE Processed Dataset	Original Dataset	CLAHE Processed Dataset
Mean	394.36	901.58	0.88	0.84	0.03	0.01	0.18	0.10
Standard Deviation	218.76	337.31	0.07	0.06	0.01	0.01	0.06	0.03

values more consistently across the image, which improves visibility of both darker and brighter regions.

Second, we performed a texture analysis using the GLCM to evaluate key textural features such as contrast, correlation, energy, and homogeneity. A comparative study of GLCM features for the original and CLAHE-enhanced datasets across 21 land use classes from the UC-

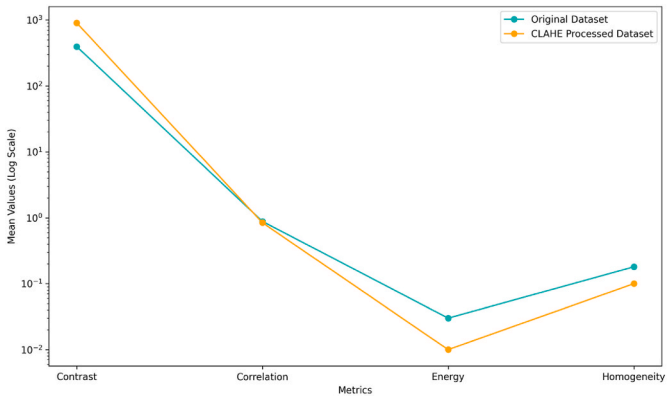

Fig. 7. A comparison of log-scaled mean GLCM features (contrast, correlation, energy, and homogeneity) scores across 21 classes between the original dataset and the CLAHE processed dataset.

Table 8
Classification performance of three CNN models on the UC-Merced Land Use Dataset.

Model	Augmentation	CLAHE Processing	Accuracy
Model A	No	No	92.14
Model B	Yes	No	96.43
Model C	Yes	Yes	97.62

Merced Land Use Dataset is shown in Table 6. The mean contrast rose significantly from 394.36 to 901.58, as can be seen in Table 7, indicating that CLAHE has increased the difference in intensity between neighboring pixels. The increased contrast sharpens edges and improves visibility of details, which is especially helpful in remote sensing applications where precise land use classification depends on the ability to distinguish between subtle features like individual houses, roads, or vegetation types. The mean correlation reduced marginally from 0.88 to 0.84, indicating a small reduction in linear reliance between pixel intensities. Because correlation represents the consistency of pixel associations, this drop implies that CLAHE provided a little degree of variability, breaking up some highly uniform regions. This minor increase in pixel diversity can help to avoid excessive smoothing and preserve subtle texture variances, which are critical for distinguishing across classes with similar tones or textures. There is also a decline in textural uniformity, as evidenced by the mean energy dropping from 0.03 to 0.01. Decreased energy implies more intricate textures and larger intensity variations in the CLAHE-processed images. This change aligns with CLAHE's objective of improving contrast by distributing intensity levels more evenly, reducing large areas of uniform brightness or darkness, and enriching the image's textural diversity. As a result, this additional detail may increase model performance by providing inputs with richer information. Last but not least, the mean homogeneity dropped from 0.18 to 0.10, indicating smoother transitions and less uniformity throughout the image. Although this decrease may result in some noise, it is a normal byproduct of the contrast enhancement. The images now have more sudden intensity variations due to lower homogeneity, which may make it easier to distinguish intricate textures within each class. Fig. 7 shows the comparison of log-scaled mean GLCM features between the original dataset and the CLAHE-processed dataset. Where it is evident that the contrast of the CLAHE-processed dataset is higher than that of the original dataset, while correlation, energy, and homogeneity have somewhat decreased. The results align with the study (Gadkari, 2004), which demonstrated that contrast and entropy consistently increased as image quality increased, whereas energy and homogeneity decreased. All things considered, these changes imply that CLAHE has improved the UC-Merced dataset's detail, contrast, and textural diversity, yielding features that are easier to see and differentiate. By providing high-contrast inputs, this could increase the classification accuracy of remote sensing models.

The three models' results suggest that data augmentation and CLAHE processing improve classification accuracy on the UC-Merced dataset, as shown in Table 8. Model A, trained on the original dataset with no enhancements, obtained an accuracy of 92.14 %. This baseline result indicates that the model performs pretty well on original data, but it lacks the robustness required for fine-grained classification. Model B achieved a significant improvement in accuracy to 96.43 % by using data augmentation. The 4 % improvement over Model A demonstrates that simply exposing the model to different transformations of the training images through data augmentation improves the model's ability for

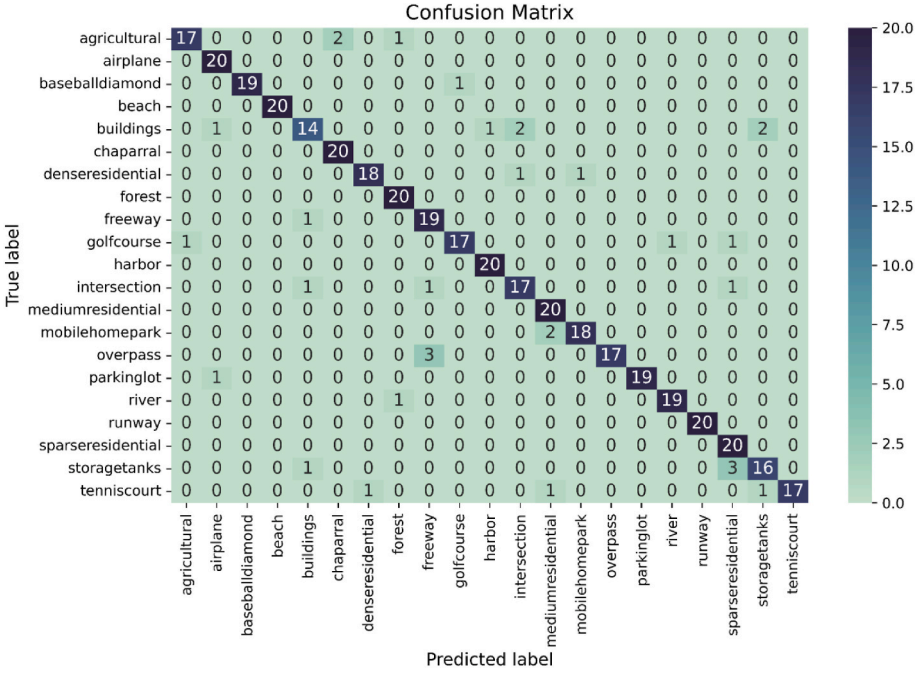

Fig. 8. Confusion matrix of Model A illustrating classification performance across 21 land use classes in the UC-Merced Land Use Dataset.

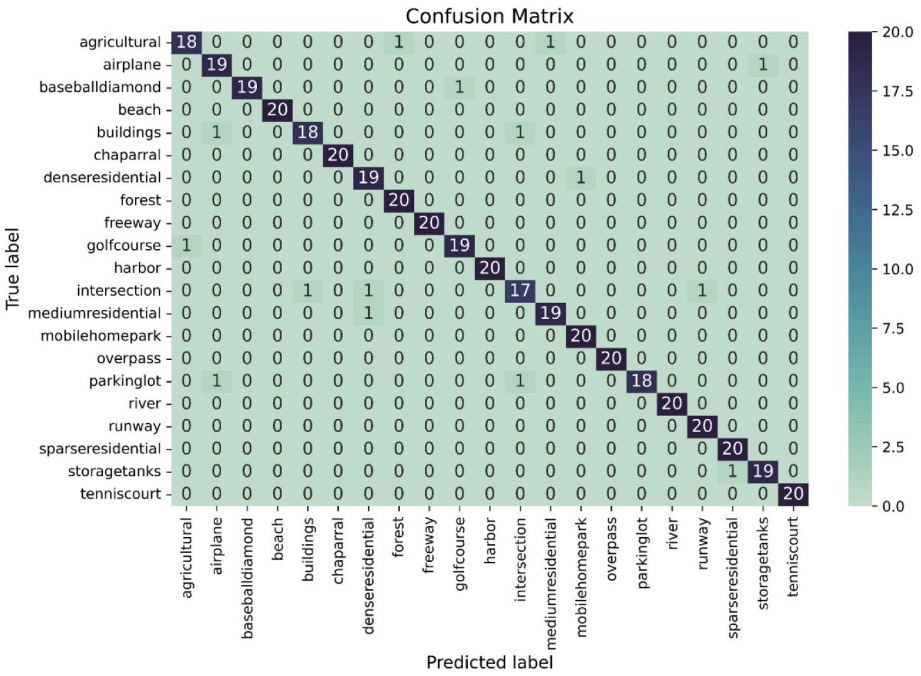

Fig. 9. Confusion matrix of Model B illustrating classification performance across 21 land use classes in the UC-Merced Land Use Dataset.

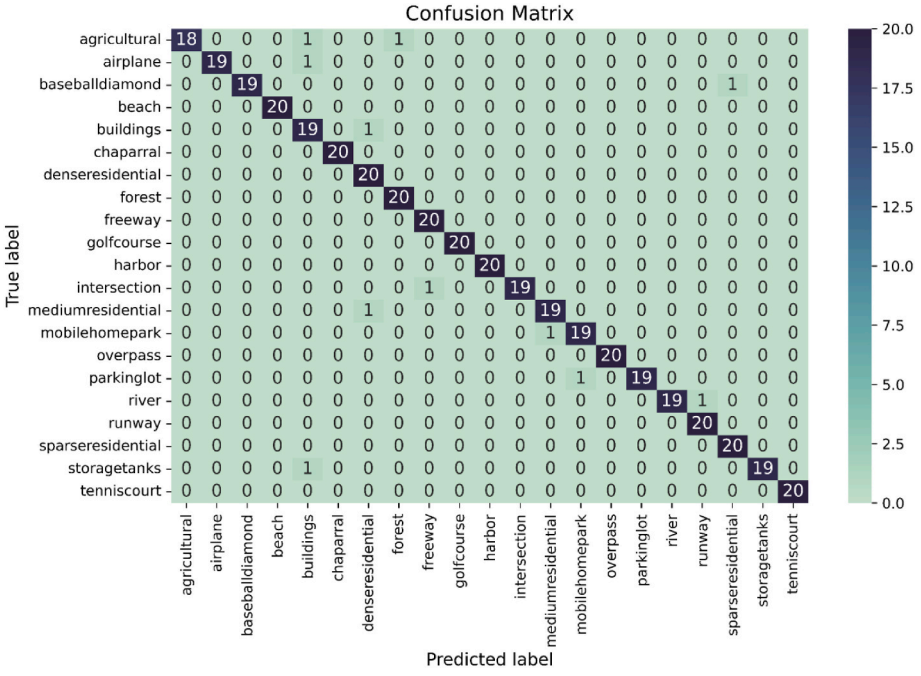

Fig. 10. Confusion matrix of Model C illustrating classification performance across 21 land use classes in the UC-Merced Land Use Dataset.

generalization. This diversity in training data helps in enhancing the model's ability to adapt to changes in real-world scenarios, leading to improved accuracy. Model C, trained with both data augmentation and CLAHE processing, achieved the best accuracy of 97.62 %. The addition of CLAHE alongside augmentation further improved accuracy by 1.19 % compared to Model B. CLAHE's role in enhancing contrast and highlighting finer details appears to complement data augmentation, providing a richer set of features for the model to learn from. This combination allows the model to better distinguish between classes, resulting in the highest accuracy of the three models.

Figs. 8–10 show the confusion matrices for Models A, B, and C, respectively, detailing classification performance across 21 land use classes in the UC-Merced Land Use Dataset. Each matrix depicts the distribution of true positive rates along the diagonal, with off-diagonal cells representing misclassification. Model A, trained on the original dataset, gives a baseline classification but has higher misclassifications, particularly in related classes, due to a lack of preprocessing. Model B, trained on the augmented dataset, shows higher accuracy, as evidenced by a more prominent diagonal pattern and fewer misclassifications, implying that data augmentation improves generalization. Model C, trained on the dataset using both augmentation and CLAHE processing, obtains the maximum accuracy, with a sharper diagonal and fewer

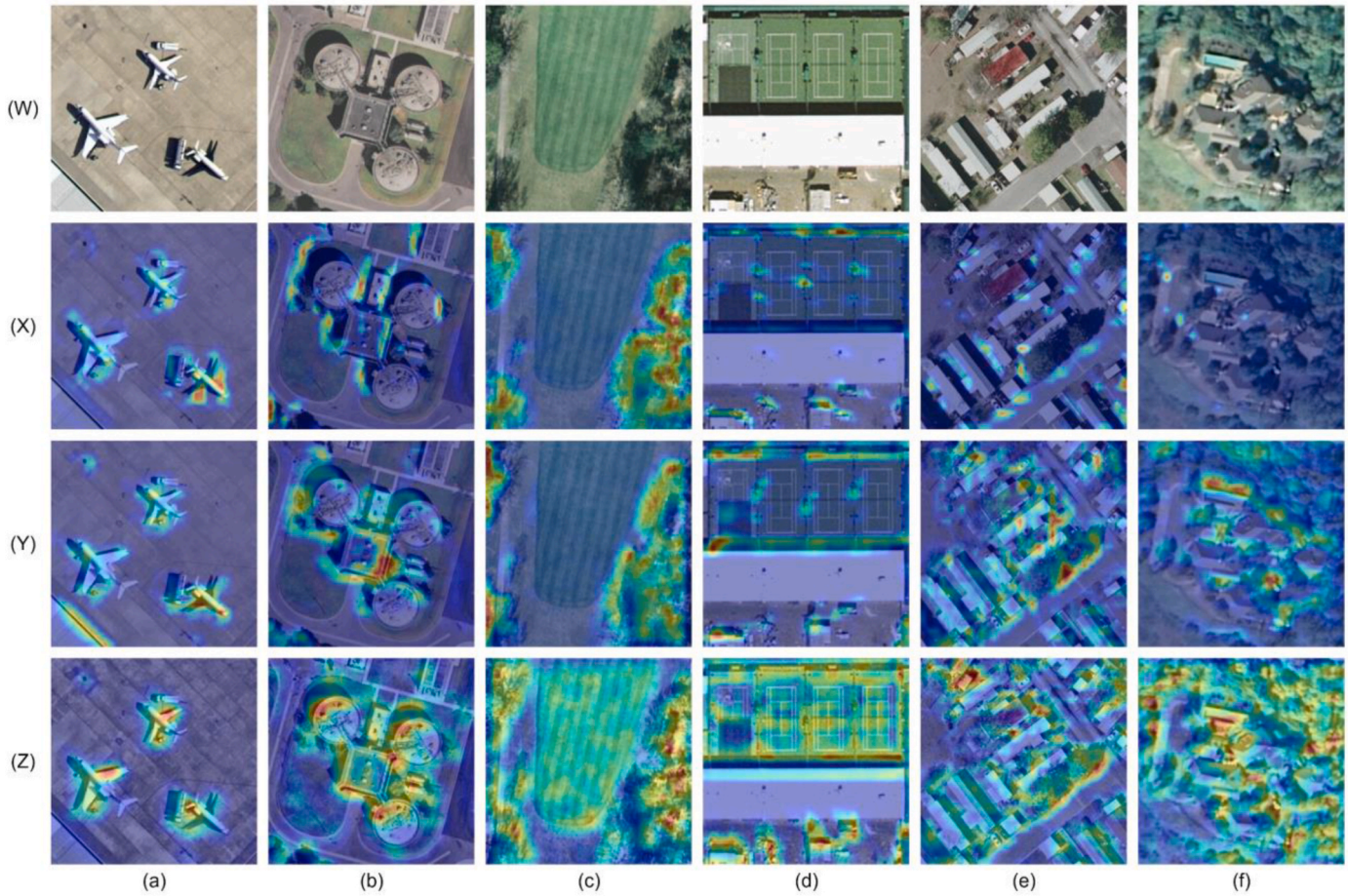

Fig. 11. Comparison of Grad-CAM visualizations generated by the same lightweight CNN architecture trained using three different approaches, resulting in three different models: Model A, Model B, and, Model C on six land use classes from the UC-Merced dataset: (a) airplane, (b) storage tanks, (c) golf course, (d) tennis court, (e) mobile home park, and (f) sparse residential. Row (W) shows the original images, while rows (X), (Y), and (Z) display Grad-CAM heatmaps overlaid on the images using three models: (X) Model A trained on the dataset without augmentation and CLAHE processing, (Y) Model B trained with augmentation but without CLAHE, and (Z) Model C trained with both augmentation and CLAHE processing.

misclassifications, indicating that CLAHE improves features and reduces class overlap.

4.4. Grad-CAM visualization

The Grad-CAM visualizations reveal a notable enhancement in feature localization and concentration from Model A to Model C, particularly when it comes to capturing class-relevant structures. As can be seen in Fig. 11, in the airplane class, Model A's activations are dispersed, omitting significant areas of the aircraft's body. Model B shows more concentrated red and yellow activation areas that cover a larger portion of the airplane body, indicating that data augmentation aids in the model's ability to more successfully concentrate on important structures. Training with both augmentation and CLAHE, Model C demonstrates the most refined attention, almost entirely covering the airplane body with red and yellow color, indicating precise feature focus and improved ability to identify and highlight critical class characteristics. This trend is also seen in other classes. For example, in the storage tanks and tennis court classes, Model A exhibits less focused activations, whereas Model B's activations become more coherent, matching the key components within each class. Although Models A and B did not focus on the course and court, which are the key distinguishing features of these classes, this suggests that they struggle to identify and prioritize essential spatial patterns specific to these categories.

Instead, their Grad-CAM heatmaps show attention spread out over unrelated areas or background features, indicating a lack of precision in feature recognition. However, in the Grad-CAM heatmaps, Model C, which used both augmentation and CLAHE processing, generated red and yellow activation on the course and court areas in the Grad-CAM heatmaps. The increased attention given to primary areas indicates that CLAHE processing has enhanced the model's capacity for detecting minor but essential structural components in the images. The occurrence of red and yellow tones in the court and course sections of Model C's heatmaps indicates a high level of confidence in these regions, underscoring their significance for class identification. The improved feature localization is likely because of CLAHE's impact on contrast enhancement, which sharpens borders and textures, enabling the model to more effectively distinguish and prioritize important regions, hence leading to increased classification accuracy. Model C consistently has the highest intensity and concentrated red and yellow activations around significant objects, such as golf course boundaries and tennis court lines, yielding the most accurate feature localization. Model C's enhanced emphasis indicates that CLAHE processing increased contrast and texture visibility, enabling the model to more accurately differentiate distinct areas relevant to each class while reducing background noise. The improvement of feature localization in the models highlights the combined advantages of augmentation and CLAHE processing. Although augmentation alone improves focus and generalization, CLAHE further sharpens the model's attention to important details, allowing it to more confidently and accurately distinguish between similar features within classes. These results highlight how preprocessing techniques can improve both visual interpretability and classification performance,

Table 9
Comparison of the proposed method with existing approaches for UC-Merced Land Use classification.

Method	Accuracy (%)	Parameters (Million)	GFLOPS	Size (MB)
GhostNet (Shen et al., 2023)	96.19	~1.73	0.00258	5.7
MSDFF (Xue et al., 2020)	**99.76**	~134.44	~15.60	~512.87
Deep Feature Fusion (Wang et al., 2020)	98.04 ± 0.89	~10.07	~1.51	~38.41
ESPA-MSDWNet (Bai et al., 2021)	98.76 ± 0.08 (50 % training set)	2.4	**0.338**	~9.16
LmNet (Wan et al., 2021)	99.52 ± 0.24	~25	~4.2	~95.37
SE-MDPMNet (Zhang et al., 2019)	98.95 ± 0.12	5.17	3.27	~19.72
BiMobileNet (Yu et al., 2020)	99.03 ± 0.28	7.76	0.45	29.59
MLCMFNet (Chen et al., 2025)	**99.91 ± 0.1**	35	Not Mentioned	~133.35
RepFACNN (Shi et al., 2025)	99.33 ± 0.26	6.4	0.87	~24.41
Proposed Method	97.62	**0.05629**	0.3519	**0.234**

particularly in complex land-use scenes where precise structure recognition is critical.

4.5. Comparison with other work

A comparative analysis between the proposed method and other existing approaches used for UC-Merced land use classification is shown in Table 9. The proposed method achieves an accuracy of 97.62 % using only 56K parameters, 0.35 GFLOPs, and occupying just 0.234 MB of storage. While several competing models (e.g., MSDFF, MLCMFNet, and LmNet) report higher accuracies ranging from 99.33 % to 99.91 %, they come at a significantly higher cost in terms of model complexity, with parameter counts from 6 million to 134 million, FLOPs from 3.27 to over 15 GFLOPs, and model sizes from 24 MB up to 512 MB. In contrast, the proposed method strikes a balance between accuracy and efficiency trade-off, making it ideal for resource-constrained environments such as drones, satellites, or embedded systems. Despite being ultra-lightweight, it performs competitively with state-of-the-art models, demonstrating the effectiveness of the model's design and its suitability for real-time aerial scene classification.

Fig. 12 illustrates the efficiency and effectiveness of our proposed model, which achieves the highest accuracy of 97.62 % while using significantly fewer parameters, lower FLOPs, and minimal storage. This makes it especially well-suited for real-time and edge-based applications. Among the compared models, ShuffleNetV2 performs second best with the lowest FLOPs, while ResNet-50 ranks the lowest, exhibiting the poorest trade-offs, with the lowest accuracy and the highest computational and storage demands. The radar chart demonstrates that our proposed model offers the most balanced and efficient network

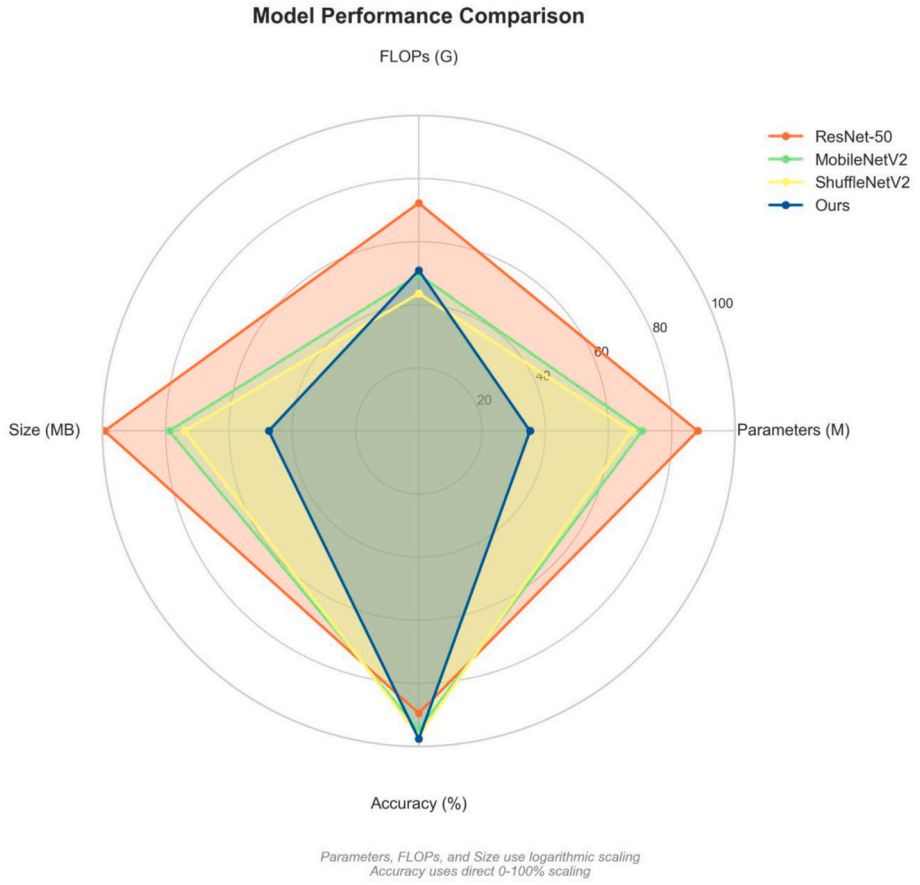

Fig. 12. Radar plot comparing ResNet-50, MobileNetV2, ShuffleNetV2, and the proposed model across four metrics: number of parameters (in millions), FLOPs (in gigaflops), model size (in megabytes), and classification accuracy (%). Higher accuracy indicates better performance, while lower values for parameters, FLOPs, and model size reflect greater computational efficiency.

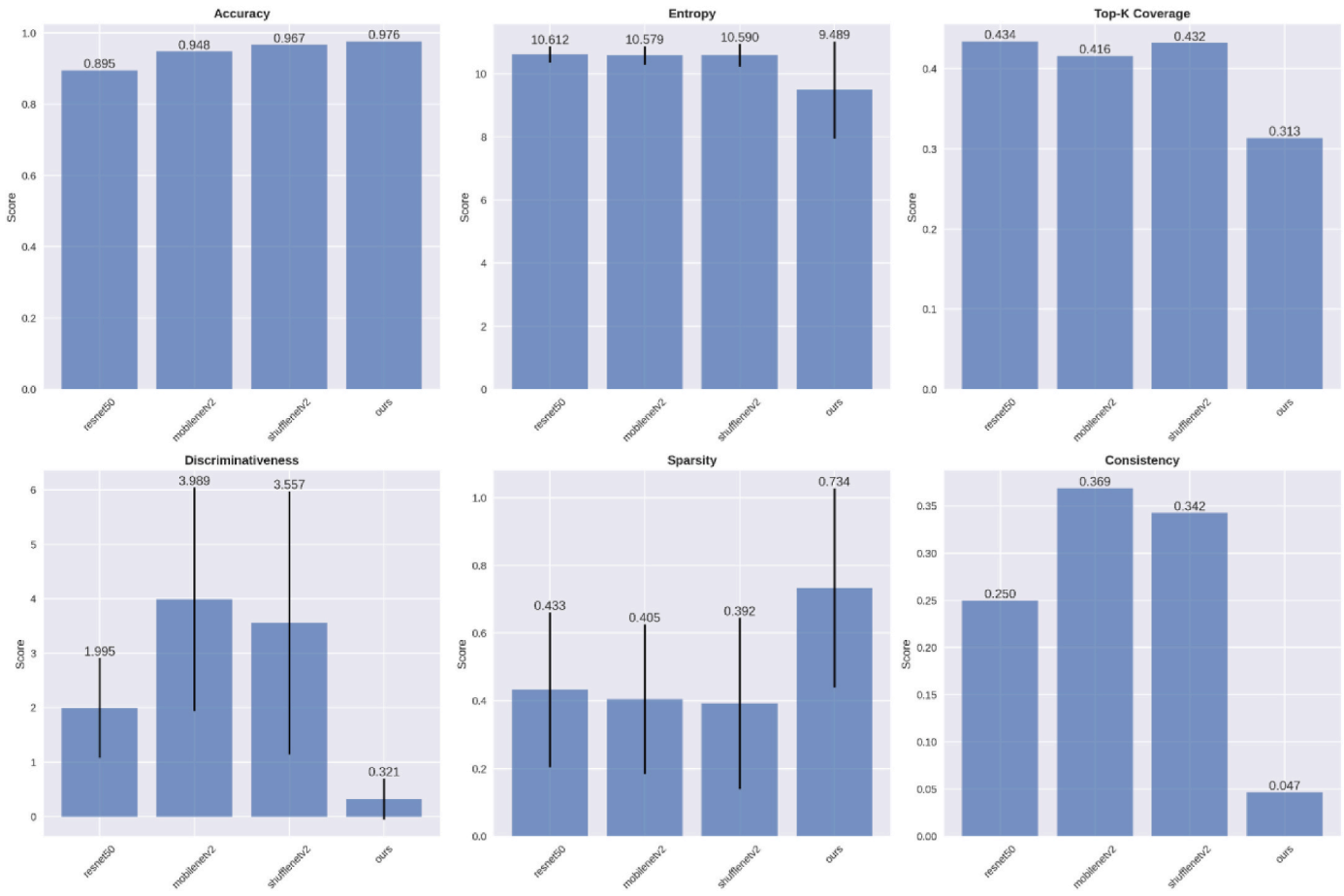

Fig. 13. Grad-CAM-based quantitative comparison of our proposed architecture with ResNet-50, MobileNetV2, and ShuffleNetV2. All models were trained under identical conditions for a fair, apple-to-apple evaluation using five metrics: Entropy, Top-K Coverage, Discriminativeness, Sparsity, and Consistency.

performance across all key metrics.

5. Discussion

Designing lightweight deep learning models for classifying many classes is challenging due to the trade-off between accuracy and parameter count. To address this for aerial scene classification, we introduced a custom Inception module that captures both local and global context, helping the network focus efficiently on key features. Additionally, we used DSC to minimize parameters and computational complexity.

As shown in Fig. 13, our proposed model achieves the lowest entropy score, 9.49, compared to ResNet-50, 10.61, indicating less dispersed attention and better focus on key scene regions, contributing to improved accuracy. In Top-K Coverage, where lower scores indicate more focused attention, our model again scores the lowest, while ResNet-50 scores the highest. This suggests our model attends to fewer but more relevant regions.

For Discriminativeness, which measures how attention varies across classes, our model achieves the lowest score of 0.32, while ResNet-50 scores the highest of 1.99. The lower value in our model implies its attention maps are more uniform across classes, likely due to shared global textures (e.g., grass in golf courses and baseball fields). Rather than relying on class-specific hotspots, our model emphasizes global scene context. In terms of Sparsity, our model scores the highest 0.734, indicating highly concentrated attention on a few critical regions, unlike ResNet-50, 0.433. However, our model exhibits lower localization consistency, meaning its attention shifts more across samples, adapting dynamically to scene variation. In contrast, MobileNetV2 and ShuffleNetV2 demonstrate broader attention due to high entropy score and higher class-specific discriminativeness, which, while effective, are less efficient for scene-level understanding. Overall, our model learns a context-aware, sparse attention mechanism that prioritizes a few highly relevant patches, making it well-suited for aerial scene classification and explaining its superior accuracy.

Although we initially used a simple train–test split, the test set was also used for validation during training. To mitigate overfitting and ensure evaluation on completely unseen data, we introduced an additional test set by applying vertical flipping and 90° rotation to the original test set. This ensures the new test data remains entirely separate from training and validation. The updated evaluation more accurately reflects the model's generalization ability. We assessed performance using both Grad-CAM-based quantitative metrics and accuracy, with results illustrated in Fig. 14. The pattern holds: our model continues to outperform larger, more complex architectures in both accuracy and context-aware attention. Despite a 2 % drop in accuracy on this unseen data, the strong performance suggests minimal risk of overfitting and confirms the model's ability to generalize effectively.

Additionally, we evaluated the impact of 7×7 kernels on performance by modifying our best-performing Model C. We removed the 7×7 kernel while keeping all other configurations unchanged. The original Model C achieved an accuracy of 97.62 %, whereas the version without the 7×7 kernel scored 96.90 %, highlighting the contribution of larger kernels in capturing broader contextual features.

Our proposed lightweight CNN model offers several managerial advantages for organizations relying on remote sensing and aerial scene analysis. Its low computational cost and minimal parameter size (only 56K) make it ideal for real-time deployment on edge devices such as

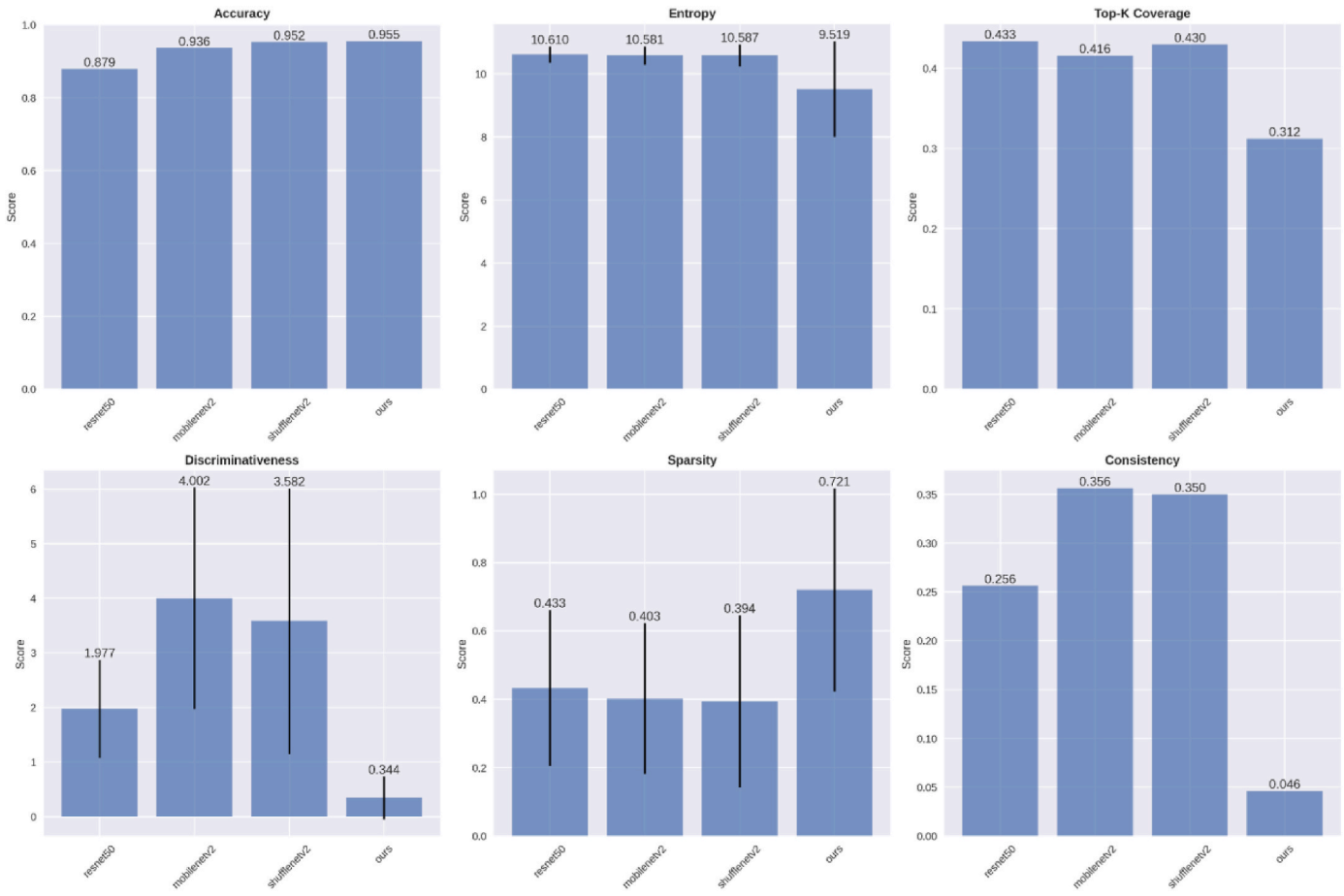

Fig. 14. Grad-CAM quantitative analysis on the augmented (unseen) test set to evaluate model generalization.

drones and satellites, significantly reducing infrastructure and energy expenses. This enables cost-effective and scalable adoption, particularly in resource-constrained settings like agriculture, urban planning, and environmental monitoring. Additionally, the integration of contrast enhancement and data augmentation techniques ensures reliable performance across diverse and low-quality image conditions, enhancing operational flexibility. The use of Grad-CAM visualizations further supports transparent decision-making by providing visual explanations of model predictions, an essential feature for fostering trust, supporting policy compliance, and facilitating human AI collaboration in critical field applications.

While the proposed lightweight CNN model demonstrates high accuracy and strong efficiency on the UC Merced dataset, there are several important limitations that must be acknowledged to contextualize its applicability and guide future work. Our evaluation is limited to a single, relatively small-scale dataset composed of 21 scene classes with uniform image resolution. As such, the model's generalizability to more diverse and large-scale remote sensing datasets such as NWPU-RESISC45, DOTA, or BigEarthNet remains untested. These datasets often include a wider range of scene complexities, scales, and noise characteristics, which could pose additional challenges for lightweight models. While we designed the model to be computationally efficient, requiring only 56K parameters, its real-world scalability has not been verified on actual edge devices such as UAVs, Raspberry Pi units, or other embedded platforms. Metrics like battery consumption, frame rate per second, thermal behavior, and memory footprint in live applications are critical for confirming true deployability. Furthermore, inference benchmarking was performed using desktop-class GPUs, which may not reflect real-time constraints faced in field environments. Although the model performs well in its current scope, future studies should explore generalization across datasets, develop integrated enhancement pipelines, validate edge-device deployment, and adopt robust explainability frameworks to fully realize the model's practical potential in real-world remote sensing applications.

6. Conclusion

6.1. Findings

This study demonstrates that the integration of targeted preprocessing techniques and multi-scale convolutional filters can significantly enhance the performance of lightweight CNN models in aerial scene classification tasks. First, CLAHE was found to be effective in improving feature visibility, particularly in scenes with low contrast or complex illumination conditions. By improving the separability of foreground and background regions, CLAHE ensures that the model receives more discriminative features, thereby reducing the dependency on deep, computationally heavy architectures. Second, the implementation of semantic-preserving data augmentation substantially improved generalization. The 5-fold augmentation strategy diversified the training set without distorting the semantic content of aerial scenes, enabling the lightweight model to learn robust representations despite its limited parameter capacity. This result indicates that even for small-scale models, effective augmentation can reduce overfitting and narrow the performance gap with larger, state-of-the-art architectures. Finally, the adoption of multi-scale convolutional filters (1×1, 3×3, 5×5, and 7×7) proved crucial in addressing the lack of global context awareness in lightweight CNNs. While small kernels (e.g., 3×3) are effective for fine-grained feature extraction, larger kernels, particularly the 7×7 filters, enabled the network to capture broader spatial dependencies that

are essential in aerial scene interpretation, such as urban layouts or agricultural patterns. This multi-scale fusion allowed the model to balance local and global feature learning while maintaining an overall lightweight architecture, ultimately achieving a competitive accuracy of 97.62 % with only 56,293 parameters.

6.2. Research limitations

Despite the promising results, this study has several limitations. First, although the proposed lightweight CNN demonstrates strong performance on the UC Merced Land Use Dataset, its generalizability to larger and more diverse datasets (e.g., high-resolution satellite imagery or cross-domain aerial datasets) remains to be validated. Second, using multi-scale filters improves performance, but larger kernels like 7×7 increase computation and may not work well on devices with limited resources like embedded systems or low-power drones. Finally, this study did not explore the integration of advanced attention mechanisms or frequency-domain methods, which could further enhance feature extraction while maintaining low computational cost.

6.3. Recommendations for future research

While 7×7 kernels proved effective for modeling global context, their computational cost constrains the design of lightweight networks. Future research should focus on efficient alternatives for large receptive field modeling, such as employing dilated convolutions to expand the receptive field without increasing kernel size or parameter count, leveraging frequency-domain approaches to complement spatial-domain convolutions for capturing global structures, incorporating attention-based mechanisms to selectively emphasize context-rich regions while reducing redundant computation, and exploring dynamic kernel selection strategies that adaptively determine kernel sizes based on scene complexity. Furthermore, optimizing CLAHE and augmentation pipelines for faster on-device processing could further strengthen the applicability of lightweight CNNs in real-world remote sensing tasks.

CRediT authorship contribution statement

Md Mahbub Hasan Rakib: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Md Yearat Hossain:** Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rashedur M. Rahman:** Writing – original draft, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Faculty Research Grant [CTRG-24-SEPS-36], North South University, Bashundhara, Dhaka 1229, Bangladesh.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.engappai.2025.112654.

Data availability

Data will be made available on request.

References

Aboshosha, Sahar, Zahran, Osama, Dessouky, Moawad I., Abd El-Samie, Fathi E., 2019. Resolution and quality enhancement of images using interpolation and contrast limited adaptive histogram equalization. Multimed. Tool. Appl. 78, 18751–18786.
Bai, Lin, Liu, Qingxin, Li, Cuiling, Zhu, Chunlin, Ye, Zhen, Xi, Meng, 2021. A lightweight and multiscale network for remote sensing image scene classification. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Bello, Olalekan Mumin, Aina, Yusuf Adedoyin, 2014. Satellite remote sensing as a tool in disaster management and sustainable development: towards a synergistic approach. Proced. Soc. Behav. Sci. 120, 365–373.
Castanyer, Roger Creus, Martínez-Fernández, Silverio, Franch, Xavier, 2021. Integration of convolutional neural networks in mobile applications. In: 2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN). IEEE, pp. 27–34.
Chen, Anzhi, Xu, Mengyang, 2025. Remote sensing image scene classification based on mutual learning with complementary multi-features. IEEE Access.
Cheng, Gong, Han, Junwei, Lu, Xiaoqiang, 2017. Remote sensing image scene classification: benchmark and state of the art. Proc. IEEE 105 (10), 1865–1883.
Chollet, François, 2017. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258.
Gadkari, Dhanashree, 2004. Image Quality Analysis Using GLCM.
Garg, Diksha, Garg, Naresh Kumar, Kumar, Munish, 2018. Underwater image enhancement using blending of CLAHE and percentile methodologies. Multimed. Tool. Appl. 77, 26545–26561.
Gui-Song, Xia, Hu, Jingwen, Hu, Fan, Shi, Baoguang, Xiang, Bai, Zhong, Yanfei, Zhang, Liangpei, Lu, Xiaoqiang, 2017. AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Rem. Sens. 55 (7), 3965–3981.
Haralick, Robert M., Shanmugam, Karthikeyan, 1973. Its' Hak Dinstein. "Textural features for image classification.". IEEE Trans. sys., Man.Cyber. 6, 610–621.
Harichandana, M., Sowmya, V., Sajithvariyar, V.V., Ramesh, Sivanpillai, 2020. Comparison of image enhancement techniques for rapid processing of post flood images. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 44, 45–50.
Huang, Yanbo, Chen, Zhong-xin, Tao, Y.U., Huang, Xiang-zhi, Gu, Xing-fa, 2018. Agricultural remote sensing big data: management and applications. J. Integr. Agric. 17 (9), 1915–1931.
Huang, Zhou, Qi, Houji, Kang, Chaogui, Su, Yuelong, Liu, Yu, 2020. An ensemble learning approach for urban land use mapping based on remote sensing imagery and social sensing data. Remote Sens. 12 (19), 3254.
Kemper, H., Kemper, G., 2020. Sensor fusion, GIS and AI technologies for disaster management. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 43, 1677–1683.
Kucharczyk, Maja, Hugenholtz, Chris H., 2021. Remote sensing of natural hazard-related disasters with small drones: global trends, biases, and research opportunities. Rem. Sens. Environ. 264, 112577.
Lin, M., 2013. Network in network. Arxiv Preprint arXiv:1312.4400.
Liu, Na, Wan, Lihong, Zhang, Yu, Zhou, Tao, Huo, Hong, Fang, Tao, 2018. Exploiting convolutional neural networks with deeply local description for remote sensing image classification. IEEE Access 6, 11215–11228.
Liu, Tao, Yang, Lexie, Lunga, Dalton, 2021. Change detection using deep learning approach with object-based image analysis. Rem. Sens. Environ. 256, 112308.
Liu, Hou-I., Galindo, Marco, Xie, Hongxia, Wong, Lai-Kuan, Shuai, Hong-Han, Li, Yung-Hui, Cheng, Wen-Huang, 2024. Lightweight deep learning for resource-constrained environments: a survey. ACM Comput. Surv. 56 (10), 1–42.
Luus, Francois PS., Salmon, Brian P., Van den Bergh, Frans, Maharaj, Bodhaswar Tikanath Jugpershad, 2015. Multiview deep learning for land-use classification. IEEE Geosci. Remote Sens. Lett. 12 (12), 2448–2452.
Malik, Rahul, Dhir, Renu, Mittal, Sudesh Kumar, 2019. Remote sensing and landsat image enhancement using multiobjective PSO based local detail enhancement. J. Ambient Intell. Hum. Comput. 10, 3563–3571.
MohanRajan, Sam Navin, Loganathan, Agilandeeswari, Manoharan, Prabukumar, 2020. Survey on Land Use/Land Cover (LU/LC) change analysis in remote sensing and GIS environment: techniques and challenges. Environ. Sci. Pollut. Control Ser. 27 (24), 29900–29926.
Padró, Joan-Cristian, Muñoz, Francisco-Javier, Planas, Jordi, Pons, Xavier, 2019. Comparison of four UAV georeferencing methods for environmental monitoring purposes focusing on the combined use with airborne and satellite remote sensing platforms. International journal of applied earth observation and geoinformation 75, 130–140.
Pietrołaj, Mariusz, Blok, Marek, 2024. Resource constrained neural network training. Sci. Rep. 14 (1), 2421.
Precedence Research. "Aerial Imaging Market Size to Hit USD 16.46 Billion by 2034." Precedence Research, accessed July–August 2025.
Salem, Nema, Malik, Hebatullah, Shams, Asmaa, 2019. Medical image enhancement based on histogram algorithms. Procedia Comput. Sci. 163, 300–311.
Santos, dos, Marciano, Jucelino Cardoso, Arantes Carrijo, Gilberto, , Cristiane de Fátima dos Santos Cardoso, Ferreira, Júlio César, Sousa, Pedro Moises, Patrocínio, Ana Cláudia, 2020. Fundus image quality enhancement for blood vessel detection via a neural network using CLAHE and Wiener filter. Res. Biomed. Eng. 36, 107–119.
Selvaraju, Ramprasaath R., Cogswell, Michael, Das, Abhishek, Vedantam, Ramakrishna, Devi, Parikh, Batra, Dhruv, 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.
Shawky, Osama A., Ahmed, Hagag, El-Dahshan, El-Sayed A., Ismail, Manal A., 2020. Remote sensing image scene classification using CNN-MLP with data augmentation. Optik 221, 165356.

Shen, Xiaole, Wang, Hongfeng, Wei, Biyun, Cao, Jinzhou, 2023. Real-time scene classification of unmanned aerial vehicles remote sensing image based on modified GhostNet. PLoS One 18 (6), e0286873.

Shi, Cuiping, Ding, Mengxiang, Wang, Liguo, 2025. Re-Parameterized feature aggregation convolutional neural network for remote sensing scene image classification. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vincent, Vanhoucke, Rabinovich, Andrew, 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Tabani, Hamid, Balasubramaniam, Ajay, Arani, Elahe, Zonooz, Bahram, 2021. Challenges and Obstacles Towards Deploying Deep Learning Models on Mobile Devices arXiv preprint arXiv:2105.02613.

Tuia, Devis, Pasolli, E., Emery, William J., 2011. Using active learning to adapt remote sensing image classifiers. Rem. Sens. Environ. 115 (9), 2232–2242.

Vidhya, Ganesh R., Ramesh, H., 2017. Effectiveness of contrast limited adaptive histogram equalization technique on multispectral satellite imagery. In: Proceedings of the International Conference on Video and Image Processing, pp. 234–239.

Wan, Huiyao, Chen, Jie, Huang, Zhixiang, Feng, Yun, Zhou, Zheng, Liu, Xiaoping, Yao, Baidong, Xu, Tao, 2021. Lightweight channel attention and multiscale feature fusion discrimination for remote sensing scene classification. IEEE Access 9, 94586–94600.

Wang, Heng, Yu, Yunlong, 2020. Deep feature fusion for high-resolution aerial scene classification. Neural Process. Lett. 51 (1), 853–865.

Weiss, Marie, Jacob, Frédéric, Duveiller, Grgory, 2020. Remote sensing for agricultural applications: a meta-review. Rem. Sens. Environ. 236, 111402.

Wellmann, Thilo, Lausch, Angela, Andersson, Erik, Knapp, Sonja, Cortinovis, Chiara, Jache, Jessica, Scheuer, Sebastian, et al., 2020. Remote sensing in urban planning: contributions towards ecologically sound policies? Landsc. Urban Plann. 204, 103921.

Xue, Wei, Dai, Xiangyang, Liu, Li, 2020. Remote sensing scene classification based on multi-structure deep features fusion. IEEE Access 8, 28746–28755.

Yi Yang and Shawn Newsam. UC merced land use dataset. Vision and Learning Lab, University of California, Merced. Accessed November 14, 2024. http://weegee.vision.ucmerced.edu/datasets/landuse.html.

Yu, Donghang, Xu, Qing, Guo, Haitao, Zhao, Chuan, Lin, Yuzhun, Li, Daoji, 2020. An efficient and lightweight convolutional neural network for remote sensing image scene classification. Sensors 20 (7), 1999.

Yuan, Qiangqiang, Shen, Huanfeng, Li, Tongwen, Li, Zhiwei, Li, Shuwen, Jiang, Yun, Xu, Hongzhang, et al., 2020. Deep learning in environmental remote sensing: achievements and challenges. Rem. Sens. Environ. 241, 111716.

Zhang, Fan, Du, Bo, Zhang, Liangpei, 2015. Scene classification via a gradient boosting random convolutional network framework. IEEE Trans. Geosci. Rem. Sens. 54 (3), 1793–1802.

Zhang, Hui, Gong, Maoguo, Zhang, Puzhao, Su, Linzhi, Shi, Jiao, 2016. Feature-level change detection using deep representation and feature change analysis for multispectral imagery. IEEE Geosci. Remote Sens. Lett. 13 (11), 1666–1670.

Zhang, Bin, Zhang, Yongjun, Wang, Shugen, 2019. A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 12 (8), 2636–2653.

Zhao, Ji, Zhong, Yanfei, Shu, Hong, Zhang, Liangpei, 2016. High-resolution image classification integrating spectral-spatial-location cues by conditional random fields. IEEE Trans. Image Process. 25 (9), 4033–4045.