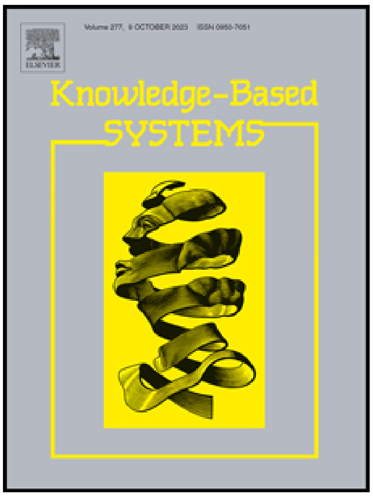


# Enhancing Land Cover Classification via Deep Ensemble Network

Muhammad Fayaz , L Minh Dang , Hyeonjoon Moon

PII: S0950-7051(24)01245-0
DOI: https://doi.org/10.1016/j.knosys.2024.112611
Reference: KNOSYS 112611

To appear in: *Knowledge-Based Systems*

Received date: 22 July 2024
Revised date: 4 October 2024
Accepted date: 6 October 2024

Please cite this article as: Muhammad Fayaz , L Minh Dang , Hyeonjoon Moon , Enhancing Land Cover Classification via Deep Ensemble Network, *Knowledge-Based Systems* (2024), doi: https://doi.org/10.1016/j.knosys.2024.112611

# Enhancing Land Cover Classification via Deep Ensemble Network

Muhammad Fayaz[1], L Minh Dang[2], and Hyeonjoon Moon[1*]

1 Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea

2 Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul, Republic of Korea

[*]Corresponding author: Hyeonjoon Moon

**Abstract**

The rapid adoption of drones has transformed industries such as agriculture, environmental monitoring, surveillance, and disaster management by enabling more efficient data collection and analysis. However, existing UAV image scene classification techniques face limitations, particularly in handling dynamic scenes, varying environmental conditions, and accurately identifying small or partially obscured objects. These challenges necessitate more advanced and robust methods. In response, this study explores ensemble learning (EL) as a powerful alternative to traditional machine learning approaches. By integrating predictions from multiple models, EL enhances accuracy, precision, and robustness in UAV-based land use and land cover classification. This research introduces a two-phase approach combining data preprocessing with feature extraction using three advanced ensemble models DenseNet201, EfficientNetV2S, and Xception employing transfer learning. These models were selected based on their top performance during preliminary evaluations. Furthermore, a soft attention mechanism is incorporated into the ensembled network to optimize feature selection, resulting in improved classification outcomes. The proposed model achieved an accuracy of 97%, precision of 96%, recall of 96%, and an F1-score of 97% on UAV image datasets. Comparative analysis reveals a 4.2% accuracy improvement with the ensembled models and a 1% boost with the advanced hybrid models. This work significantly advances UAV image scene classification, offering a practical solution to enhance decision-making precision in various applications. The ensemble system demonstrates its effectiveness in remote sensing applications, especially in land cover analysis across diverse geographical and environmental settings.

**Index Terms:** Attention mechanism, land area images, land use classification, land cover classification, Ensemble Learning, Satellite imagery, Remote Sensing,

## 1. Introduction

The widespread use of Unmanned Aerial Vehicles (UAVs), also called drones has increased in recent years and has revolutionized various industries, including agriculture, environmental monitoring, surveillance, and disaster management [1]. These versatile devices are equipped with robust imaging sensors that capture high-resolution images of vast and dynamic landscapes, making them invaluable tools for various applications [2]. One of the critical tasks in UAV-based applications is image scene classification, which involves categorizing images into predefined classes [3]. Accurate scene classification is pivotal for tasks like object detection, land-use mapping, and disaster response planning. However, classifying scenes from UAV images poses a distinct set of challenges. These difficulties stem from factors like fluctuating lighting conditions, image blurring caused by the drone's movement, alterations in camera angles and perspectives, and the existence of small or partially hidden objects. To surmount these obstacles and elevate the precision of scene classification, machine learning methods, with a particular emphasis on EL, have emerged as a prominent solution [4]. EL is a powerful approach that combines the predictions of multiple classifiers to improve overall classification performance [5]. Various EL techniques have been explored in the context of UAV image scene classification, and the choice of the ensemble method can significantly impact the final classification accuracy [6], [7].

Several methods have been developed for land area classification; however multiple challenges are associated with previous studies [8, 9]. The previous work utilized a solo architecture for land area classification, which performed insufficiently due to the range of objects captured by UAV cameras[10]. Furthermore, these studies used shallow architectures and fed the output of convolutional layers directly to the fully connected layers without selecting optimal features, resulting in suboptimal classification performance. Despite various efforts, challenges such as handling varying environmental conditions, managing dynamic scenes, and accurately identifying small or partially hidden objects persist[11, 12]. Therefore, this work aims to develop an attention-enhanced deep ensemble learning (EL) model to identify and categorize land use and cover changes accurately. By leveraging optimized deep convolutional neural networks (CNNs), we introduce an ensemble architecture that excels in classifying land use and cover changes. The model is designed to meet critical requirements, including high accuracy for reliable differentiation between land cover types, scalability to process large datasets, robustness to handle varying conditions, and interpretability to ensure the results are understandable and actionable.

The practical significance of this approach is substantial, as accurate land cover classification is vital for effective environmental monitoring, urban planning, agriculture management, and disaster response. To further improve accuracy, we incorporate a soft attention mechanism that enhances the model's ability to focus on the most relevant features in the input data, reducing noise and improving classification precision [13, 14]. This not only boosts the model's overall performance but also makes the decision-making process more transparent, thereby supporting better land management strategies and conservation efforts. The main contribution of the work is as follows.

- We have introduced an innovative integrated deep ensemble model, which combines the strengths of multiple pre-trained deep learning algorithms. Unlike traditional single-model approaches, our method leverages the complementary advantages of three baseline models. This novel ensemble strategy significantly advances the performance in UAV image scene classification, surpassing existing methods and setting a new benchmark for classification accuracy.
- Another key novelty of our approach lies in the integration of a soft attention module, which dynamically selects the most relevant features for classification tasks. This attention mechanism enhances the model's ability to focus on critical features, leading to improved accuracy in land cover classification. This integration enables the proposed ensemble learning (EL) model to better handle dynamic and constantly changing environmental conditions, making it highly resilient and robust in real-world UAV imagery applications.
- We have conducted comprehensive experiments that demonstrate our model superior performance compared to baseline models in land area classification. Moreover, an extensive ablation study was performed to identify the most optimal configuration of the ensemble model, further validating the proposed model robustness and reliability. These results highlight the practical applicability of our model, providing new insights and setting a higher precision for UAV-based land cover classification tasks.

The rest of the work is organized as; Section 2 provides an overview of previous research in this field; Section 3 presents the methodology. The experimental Results are discussed in section 4 and finally conclude the article in Section 5.

## 2. Literature Review

These studies encompass diverse approaches, from optimizing spectral bands to employing advanced deep learning techniques to improve land cover classification accuracy and effectiveness in various remote sensing applications. Researchers continuously enhance methodologies, contributing to the advancement of environmental monitoring, wildlife habitat prediction, and other domains [15, 16]. Researchers in land cover classification employ deep learning techniques, ranging from CNN-based multispectral LiDAR systems to hybrid feature optimization models [17],[18]. For instance, Pan et al. [19] introduced Multispectral LiDAR land cover classification using a Convolutional Neural Network, optimizing parameters for improved performance. Similarly, Kwan et al. [20] explored the effectiveness of using CNN-based models for classifying land cover using different combinations of bands, including RGB, NIR, and LiDAR data.

Moreover, Zhang et al. [21] presented the MLCG-OCNN algorithm, focusing on object discrimination and spectral pattern learning, further refining land cover classification techniques.

Rajendran et al. [22] adopted a hybrid feature optimization model combined with DL classifiers to achieve notable improvements in land use and land cover (LULC) classification. Additionally, Chatterjee et al. [23] proposed an unsupervised clustering method for polarimetric Synthetic Aperture Radar (SAR) images, exploring fully convolutional networks for precise land cover detection. Other researchers [24-28] have evaluated the precision of land cover classification using neural network methods, focusing on high-resolution satellite imagery, while [29] investigated multimodal DNNs for land cover classification, considering processing efficiency and network traffic. Lei Song et al. [30] and Jing chen et al. [31] employed advanced methods, including a bi-branch fusion network combining CNNs and axial cross- attention mechanisms, to enhance land cover change detection accuracy in remote sensing images. Their approach effectively integrates local and global feature extraction, leading to significant improvements in detection performance. Recently, EL methods achieved better performance than solo models.

Ensemble learning in UAV image classification has gained substantial attention, and several methods have been developed. For instance, Rahee Walambe et al. [32],[33] explore the application of EL methods, such as bagging and boosting, highlighting their ability to improve UAV image classification accuracy in challenging environmental conditions. Jin and Xu et al. [15, 34-37] demonstrated that the Gaussian process regression and ensemble models, particularly combining a Hodrick Prescott filter and neural network, can significantly improve prediction accuracy compared to individual models, underscoring the potential benefits of ensemble approaches in complex prediction tasks. Bolin Fu and Lei et al. [38], [39] focused on the stacking ensemble approaches, effectively combining DL models with traditional machine learning algorithms for real-time UAV image classification. The study highlights the effectiveness of stacking ensembles in handling diverse and dynamic scenes, making it a valuable approach for practical applications.

Colkesen et al. [40], [41] comprehensively analyze various EL techniques. These authors provide a comprehensive analysis of the strengths and weaknesses of each method, guiding the selection of the most suitable ensemble approach for specific scenarios. McCoy et al. [42] explored deep EL techniques for multimodal UAV image classification. The authors employ deep neural networks and ensemble methods to classify images from multiple sensor modalities, emphasizing the importance of leveraging diverse data sources for improved classification accuracy. Additionally, Namoun et al. [43] and Sefrin et al. [44] extend the application of EL to land cover change detection in UAV imagery. These authors propose an ensemble approach to detect and classify changes in land cover over time, showcasing the adaptability of ensemble methods to evolving environmental conditions. As evident from the literature, EL has become a pivotal component of UAV image classification, offering improved accuracy, robustness, and adaptability to various environmental conditions. These related works contribute to the growing body of knowledge in the field and showcase the diverse applications of ensemble techniques in the context of UAV and UAV imagery analysis. Therefore, we introduce an attention-enhanced ensemble learning model that meets critical requirements like higher accuracy, reliable differentiation between different land cover types, and scalability for processing large datasets. We specifically developed this model to identify land cover changes effectively, addressing the challenges associated with evolving environmental conditions and the need for precise, large-scale analysis. The advancements presented in our model underscore the growing importance of ensemble techniques in UAV imagery analysis and their pivotal role in

**Table 1:** Details of dataset, model, and performance of different methods developed for land area classification.

| Reference | Method | Dataset | Models | Year | Accuracy% |
|---|---|---|---|---|---|
| Pan et al [19] | CNN-based multispectral LiDAR system | HIS, VHR-RGB | Two-Stream CNN | 2018 | 94.8 |
| Kwan et al [20] | CNN-based DL models | RGB, NIR, LiDAR | -- | 2019 | 90.6 |
| Zhang et al [21] | MLCG-OCNN algorithm | -- | Object discrimination | | 83.45 |
| Rajendran et al. [22] | Hybrid model | | DL classifiers | 2021 | 91.2 |
| Chatterjee et al [23] | Unsupervised clustering | Polarimetric SAR images | Fully convolutional networks | 2022 | 89.7 |
| Moon et al [24] | -- | KOMPSAT-3 satellite imagery | SVM, ANN, and DNN | 2020 | 92.0 |
| Aspri et al [29] | -- | NCALM | CNN | 2020 | 83.6 |
| Walambe et al [32] | Ensemble | -- | -- | 2021 | -- |
| Fu et al [38] | | -- | -- | 2022 | 92.2 |
| Colkesen et al [40] | | -- | -- | 2022 | 92.8 |
| Deepan et al [45] | | -- | -- | 2021 | 93.2 |
| Naftalu et al [46] | | -- | -- | 2022 | 90.73 |
| Sefrin et al [44] | | -- | -- | 2020 | ------- |
| Xu et al [47] | | Vaihingen, | ATFM, DAL | 2023 | 90.57 |

| | | | | |
|---|---|---|---|---|
| | Postdam | | | |
| Ma et al[17] | -- | FENet | 2023 | 82.85 |
| Li et al [18] | Houston, Trento | EMFNet | 2021 | 96.1 |

advancing the field of land cover classification. Some standard approaches from the literature are further summarized in **Table 1**.

## 3. Methodology

This section describes the proposed framework for land area image scene classification. Our approach employs various stages, including pre-processing, augmentation, training, and evaluation, where the flow diagram illustrating these phases is given in **Figure 1.** The framework ensembles features extracted by different models at the feature level rather than simply combining prediction results from isolated models. This allows the model to integrate the strengths of multiple deep learning architectures into a unified representation before making the final classification. In this method, we leverage deep learning models for fine-tuning and transfer learning, optimizing critical hyper-parameters such as learning rate, activation functions, and batch size. Additionally, the attention module is designed to enhance the model's focus on the most relevant features, assigning varying levels of importance to different aspects of the data. This ensures that the model prioritizes critical information, improving classification accuracy and making the methodology both robust and effective. Each steps of the proposed model are further discussed in the subsequent section.

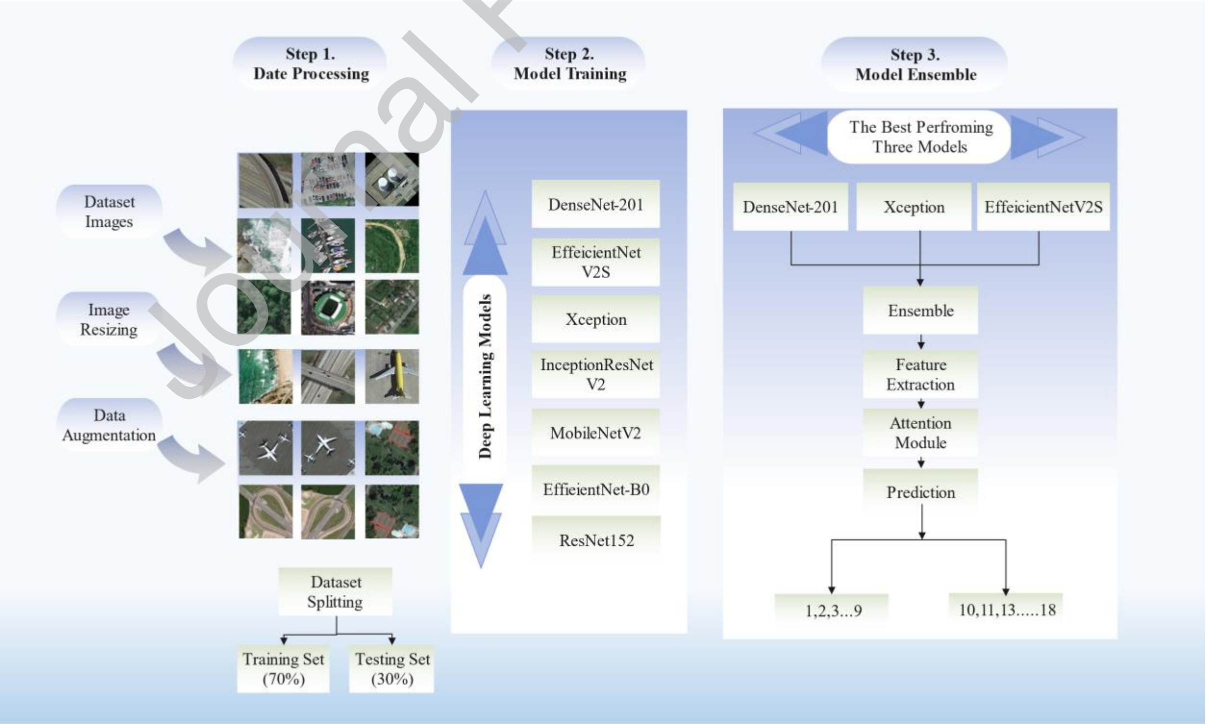


**Figure 1:** A generic diagram of the proposed model including processing, extraction of significant features, and training/classification. In processing, images are resized, and data augmentation is

applied for enhanced training accuracy. Transfer learning with DenseNet-201, EfficientNetV2S, and Xception models extract deep features. Ensemble classifiers use these features to predict class probabilities.

### 3.1. Data Augmentation:

We applied various data augmentation techniques, including rotation, flipping, and zooming, during preprocessing to prevent overfitting and enhance performance on small-scale datasets. This approach improved the model's generalization and robustness by exposing it to a broader range of scenarios, ultimately boosting its performance.

### 3.2. Data Preprocessing:

Data preparation is critical in deep learning, as it enhances input data quality, which is important for improving network performance and model generalization. In our preprocessing stage, we applied various image enhancement techniques, including rotation, flipping, scaling, and brightness adjustment, to augment the dataset and simulate diverse real-world conditions as given in Figure 2. These augmentations aimed to improve the model's ability to generalize and reduce overfitting. The preprocessing steps significantly boosted the model's accuracy and consistency, particularly in handling the diversity and complexity of UAV imagery, ensuring effective performance across different scenarios. As a first stage in our study, we used data normalization, multiplying all pixel values by 1/255 to convert them to a range of [1, -1]. This can be stated mathematically as an equation (1):

$$R_i = \frac{I - Imin}{Imax - Imin} \qquad (1)$$

Where $I$ represent the authentic data, the input maximum, minimum, and normalized data are denoted by the letters $Imax, Imin$, and $Ri$, respectively. The pixel values are scaled from their original range of [0,255] to a new range of [0,1] by simply normalizing them by 1/255. To adjust pixel values to a range of [1, -1], we use the following formula for this approach.

$$\dot{R}_i = 2\,.\left(\frac{I}{255}\right) = \frac{2I}{255} - 1$$

This transformation adjusts the normalized pixel values from the range [0, 1] to the range [-1, 1]. By including this explanation, we clarify that the 1/255 normalization is actually a specific implementation of the more general Min-Max normalization method, adapted for the standard range of image pixel values. By guaranteeing that the input data is properly scaled, normalization plays a vital role in enhancing the model's performance, especially in relation to convergence and generalization. Images were scaled to 224 × 224 × 3 before training. Data augmentation techniques included zooming (0.2), nearest complete mode setting, 90-degree rotation, and flipping, enhancing dataset variations for improved model generalization and reduced overfitting.

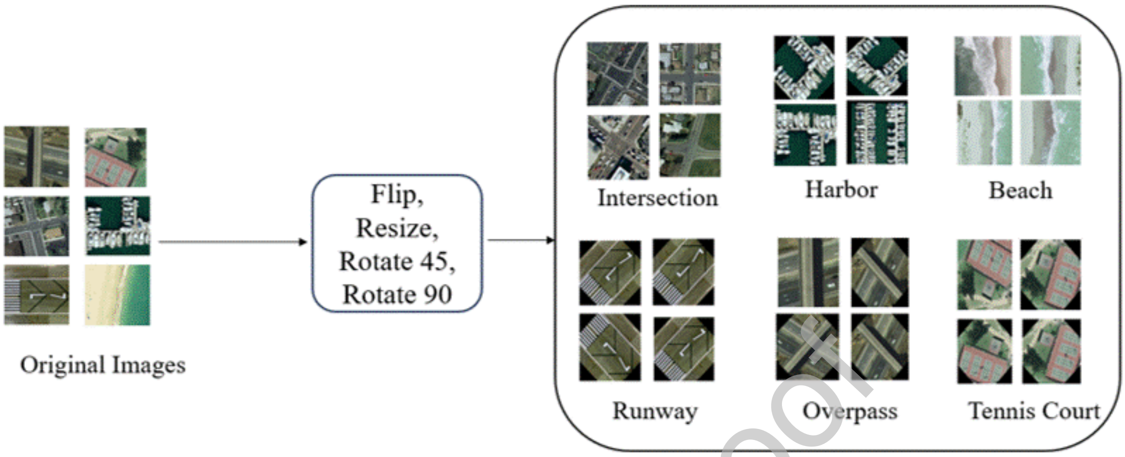


**Figure 2:** Data augmentation, including rotating, flipping, cropping, and zooming, expands the training dataset, enhancing model robustness and generalization.

### 3.3. Deep Ensemble Strategy

In the realm of UAV image scene classification, our research adopts a novel deep ensemble strategy to

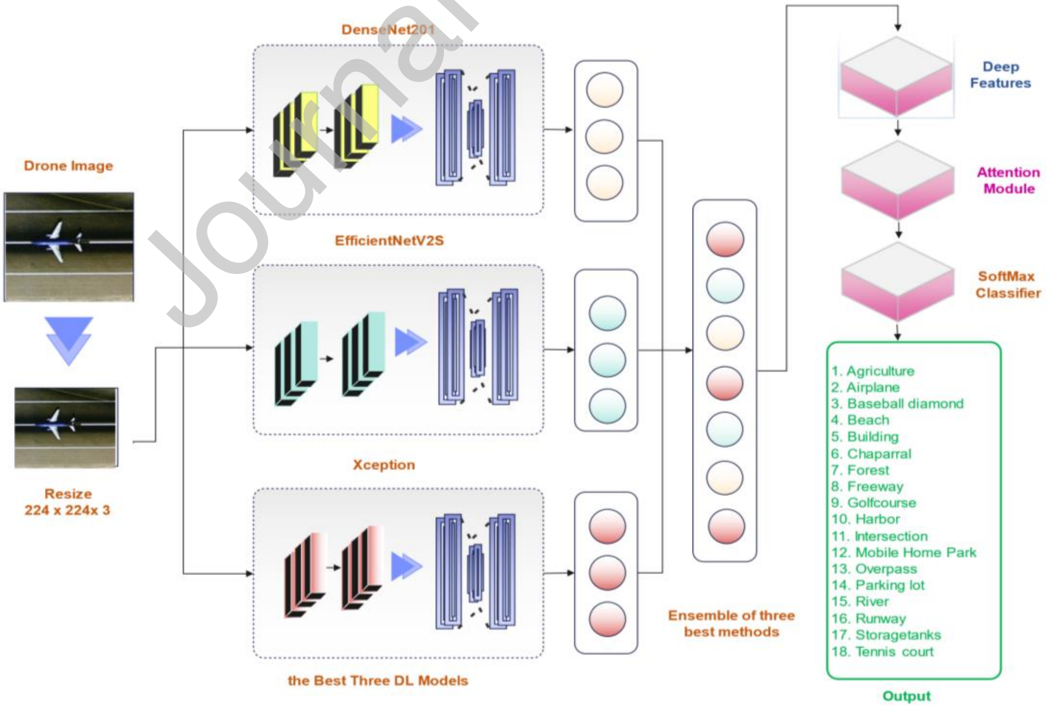


**Figure 3:** A visual depiction showcasing the deep ensemble strategy, featuring the three most successful fine-tuned models.

enhance classification accuracy. While traditional ensemble methods, such as bagging, focus on aggregating the predictions of multiple models, our approach diverges by emphasizing the ensemble of features. Specifically, we train multiple models on different subsets of features, and rather than simply aggregating predictions, we combine the learned feature representations from these models to build a more robust and informative classifier. This feature-based ensemble allows for a more diverse and comprehensive understanding of the data, which in turn contributes to improved classification outcomes.

The architecture of our ensemble method is visually depicted in **Figure 3**, which showcases the combination of three high-performing DL models: DenseNet201, EfficientNetV2S, and Xception. These models were considered for their exceptional performance in the context of UAV image scene classification. The ensemble combines multiple deep learning models, including DenseNet201, Xception, and EfficientNetV2S, among others, to leverage the strengths of each architecture. DenseNet201 was chosen for its dense connectivity, which enhances information flow and is particularly effective for high-resolution drone imagery where capturing fine details is crucial. EfficientNetV2S was selected for its balance between accuracy and computational efficiency, making it ideal for processing large UAV datasets without compromising performance. Xception, with its depth-wise separable convolutions, excels at capturing spatial hierarchies, which is essential for distinguishing complex land cover patterns.

The theoretical basis for our ensemble method lies in the principle of model complementarities, where diverse models are likely to capture different aspects of the data, thereby improving overall generalization. To optimize the ensemble, we use an averaging method combined with a soft attention mechanism. This attention mechanism dynamically adjusts the weights of model outputs based on their contribution to the final prediction, thereby ensuring that the ensemble model adapts to the varying complexity of input data. More importantly, the integration of attention mechanisms further refines the ensemble by focusing on the most relevant features, thus enhancing the predictive accuracy. This strategy is supported by theoretical insights into ensemble learning, which demonstrate that well-constructed ensembles often outperform single models, particularly in complex tasks such as image classification. To further support our choices, we conducted comparative experiments showing that these models outperformed others in terms of accuracy, precision, and robustness in land cover classification.

The mathematical framework of deep EL, as outlined in **Equation (2),** highlights our approach. Our research aims to fetch the positive aspects of layered ensemble techniques to the challenging domain of UAV image scene classification to increase the precision, accuracy, and dependability of automated scene recognition in UAV imagery.

$$f(X) = \sum_{i=1}^{n} w_z f_i(X) \tag{2}$$

Here, the collective output of all models together is represented by $f(X)$, whereas, $X$ represents the input vector, and the weight given to the $i^{th}$ model output of the $n$ algorithm is represented by $Wz$, and $fi(X)$ is the output of the $i^{th}$ model. Prediction confidence is determined by standard error predictions, as follows:

$$\sigma_e = \left\{\frac{1}{n-1}\sum\nolimits_{b=1}^{n}\left[y\left(x_j w^b\right) - y\left(x_j\right)\right]^2\right\}^{\frac{1}{2}} \tag{3}$$

Where $b(aj) \;=\; \sum k = 1\; b\;(aj;\; ck)/n$ is the predicted output for input $aj$, $n$ is the number of neural networks used, $y(aj;\; CB)$ is the predicted output for input $aj$ using the $b - th$ neural network, and a smaller $\sigma e$ indicates a more reliable model prediction. We enhance DL model performance by freezing specific layers during fine-tuning and transfer learning. Images are adjusted to 224×224×3 for feature extraction.

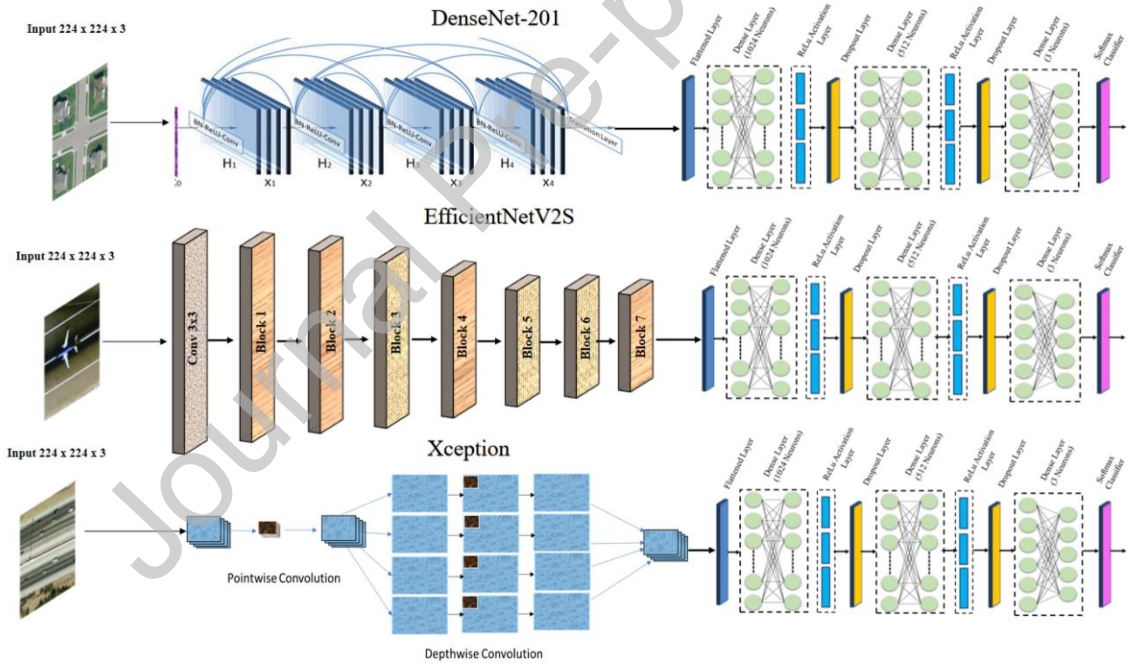


**Figure 4:** Different models ensembled in the proposed model.

We used different models such as VGG-16 [48], ResNet152, MobileNetV2 [49], InceptionV3 [48], Xception [50], Efficient Net [51], and Dense Net [52] in the model. Every model is optimized with modifications to their last layers, and finally, the nest performance is achieved in Xception, EfficientNetV2, and DenseNet. The last layer includes a dimension reduction layer, two dropout layers, and three densely connected layers, along with a SoftMax classifier in the output layer. The flattener transforms characteristics to a 1D vector, which is afterward processed through dense layers with 512 and 1024 hidden units, respectively. **Figure 4** illustrates our suggested layers. This layer employs the ReLu activation function, Provided by

$$f(x) = max(0, x) \qquad (4)$$

The dense layer is triggered first, mapping each neuron output assigned a label via the ReLu function before making a prediction. With each activation map, weights and biases are multiplied to create the probability using a linear strategy in this method. In the hidden layer, a 30% dropout was applied to prevent overfitting. As our ultimate identifier, we employ the SoftMax classifier [53], as formulated in Eq.5.

$$\boldsymbol{\sigma}(\vec{j})_i = \frac{e^{j_1}}{\sum_{z=1}^{k} e^{j_i}} \qquad (5)$$

The input vector components are $Ji$, with $k$ as the number of classes and $ji$ representing the input vector.

DenseNet-201 is a convolutional neural network (CNN) model having 201 layers; each layer is linked to every other layer in a feed-forward manner. Feature maps are provided to higher layers after receiving data from lower layers in the DenseNet architecture. DenseNet offers benefits such as enhanced feature propagation, a significant reduction in parameters, promotion of feature reutilization, and mitigation of the issue of gradients disappearing. This model outperforms the compared models in terms of speed and size. The width and height of the input image dictate its resolution, while the convolutional layers establish the network depth. DenseNet201 is a prominent computer vision model advancing through deep architecture with (3 × 3) convolutional filters.

### 3.4. Soft Attention

In this study, we used seven different deep-learning models to classify land area images, including DenseNet201, EfficientNetV2S, Xception, MobileNetV2, ResNet152, Inception V3, and EfficientNetB0. After testing, we chose three models with the highest accuracies and combined them into different pairs. Firstly, we combined DenseNet201, EfficientNetV2S, then EfficientNetV2S, Xception, and then Xception, DenseNet201and lastly, DenseNet201, EfficientNetV2S, Xception to see which pair is classifying the best together. After trying these combinations of different models, we found that the combination of all three models gave us the highest accuracy. To make classification better, we employ smart features and soft attention to these ensembled models. The soft attention mechanism in our model is designed to selectively focus on the most relevant parts of the input image data, thereby enhancing the model's ability to classify complex scenes accurately. The process begins with extracting feature maps from the input images using convolutional neural networks (CNNs), which capture various aspects of the image, such as edges, textures, and patterns. Once these feature maps are obtained, a small neural network calculates attention scores for each map. These scores are then normalized using a SoftMax function to generate attention weights, determining each feature map's relative importance. The soft attention mechanism differs from intricate attention by allowing the model to assign varying levels of importance to different features simultaneously rather than focusing on a single part of the image.

These attention weights are applied to the corresponding feature maps, amplifying the significant features while downplaying the less relevant ones. The weighted feature maps are then aggregated to create a focused representation that emphasizes the most critical parts of the image. This aggregated feature map is subsequently passed through fully connected layers for classification. By directing the model's attention to the most informative regions, the soft attention mechanism improves classification accuracy and enhances the interpretability of the model's decisions. We validated the effectiveness of this approach through experiments that demonstrated a significant and impressive performance improvement, particularly in scenarios with challenging conditions such as varying lighting and complex terrain types. This design ensures the model is robust and adaptable, making it well-suited for a wide range of classification tasks.

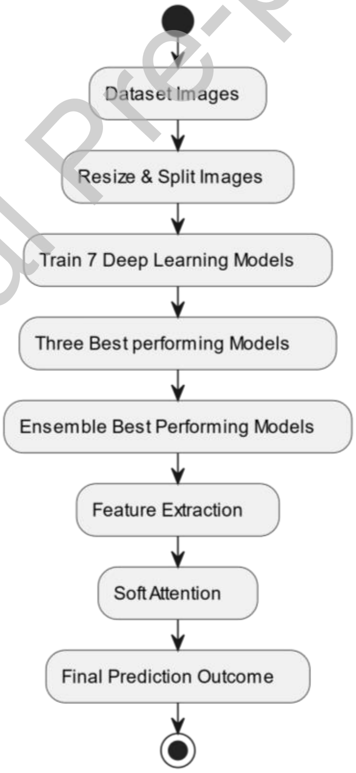


Figure 5: Flow Daigram of proposed study

Additionally, it reduces overfitting by concentrating on essential details, leading to better generalization. We used this approach to ensure that the ensemble could leverage the strengths of each model effectively, leading to improved overall accuracy.

$$f(att) = \sum_{i=1}^{n} \boldsymbol{\sigma}_i \,(X).w_z.f_i(X) \qquad (6)$$

$\sigma_i$ (x) represents the attention weight for the i-th model which is the function of input x.

### 3.5. Transfer Learning and Fine-Tuning

These steps involved in training and perfecting our models are covered in this section. First, we use pre-trained weights from the 14 million images in the ImageNet dataset that have been divided into 1000 classes. The ImageNet dataset's pre-trained weights enable faster use of earlier acquired features and enhanced image identification performance. Features unique to image classification are contained in the ImageNet weights that were obtained during training. This approach is more efficient than starting with randomly initialized weights, as it enables the model to quickly adapt to new tasks with a solid foundation of pre-learned features. Transfer learning, utilizing pre-trained weights, is more efficient and faster than using randomly initialized weights [54]. We then froze every layer of the base model to fine-tune it. Except for the last layers, which are trained using UAV images, preventing changes to the initial layers from the UCMerced_LandUse dataset. The input layers of this method retain pre-trained ImageNet weights. After training the last layers on the UCMerced_LandUse dataset, we unfreeze the entire network for further fine tuning, allowing the entire model to integrate and optimize both general and task specific features. The final models were then evaluated on test data to demonstrate the effectiveness of this transfer learning and fine-tuning approach. The flow diagram of the proposed model is given in figure 5.

## 4. Results and Discussion

This section discusses the dataset, evaluation metrics, experimental setup, and comparative analysis with supportive ablation study. Experiments were conducted on a system that runs Python 3.8, integrates the TensorFlow and Keras frameworks, and runs a 64-bit version of Windows 10, which also includes 2.80GHz processors, 24 GB of RAM with an NVIDIA GeForce RTX 3090 GPU. All the models are fine-tuned using a different set of hyperparameters, as depicted in **Table 2**. The input dimensions are 224 × 224 × 3, and a batch size of 32 is being used. Categorical Cross Entropy is applied as the loss function with an SGD optimizer. Finally, a SoftMax activation function is applied to the output layer.

### 4.1. Dataset

We obtained experimental results using a comprehensive UCMerced_LandUse image dataset [54]. A randomly selected, balanced collection of images was used to ensure accuracy, with 70% of the dataset allocated for training and the remaining 30% reserved for testing. Statistical details of the dataset are given in **Table 3**, while sample images of each class are represented in **Fig. 6**. The UCMerced_LandUse dataset consists of 18 different classes, each containing 1000 images. This comprehensive dataset allows for a robust evaluation of our models. For each class, 700 images were used for training and 300 for testing. This division ensures that the training set is sufficiently large to train the models effectively

while the test set provides a reliable measure of model performance. The detailed breakdown of the dataset is provided in **Table 3**, illustrating the balanced nature of the dataset across all classes.

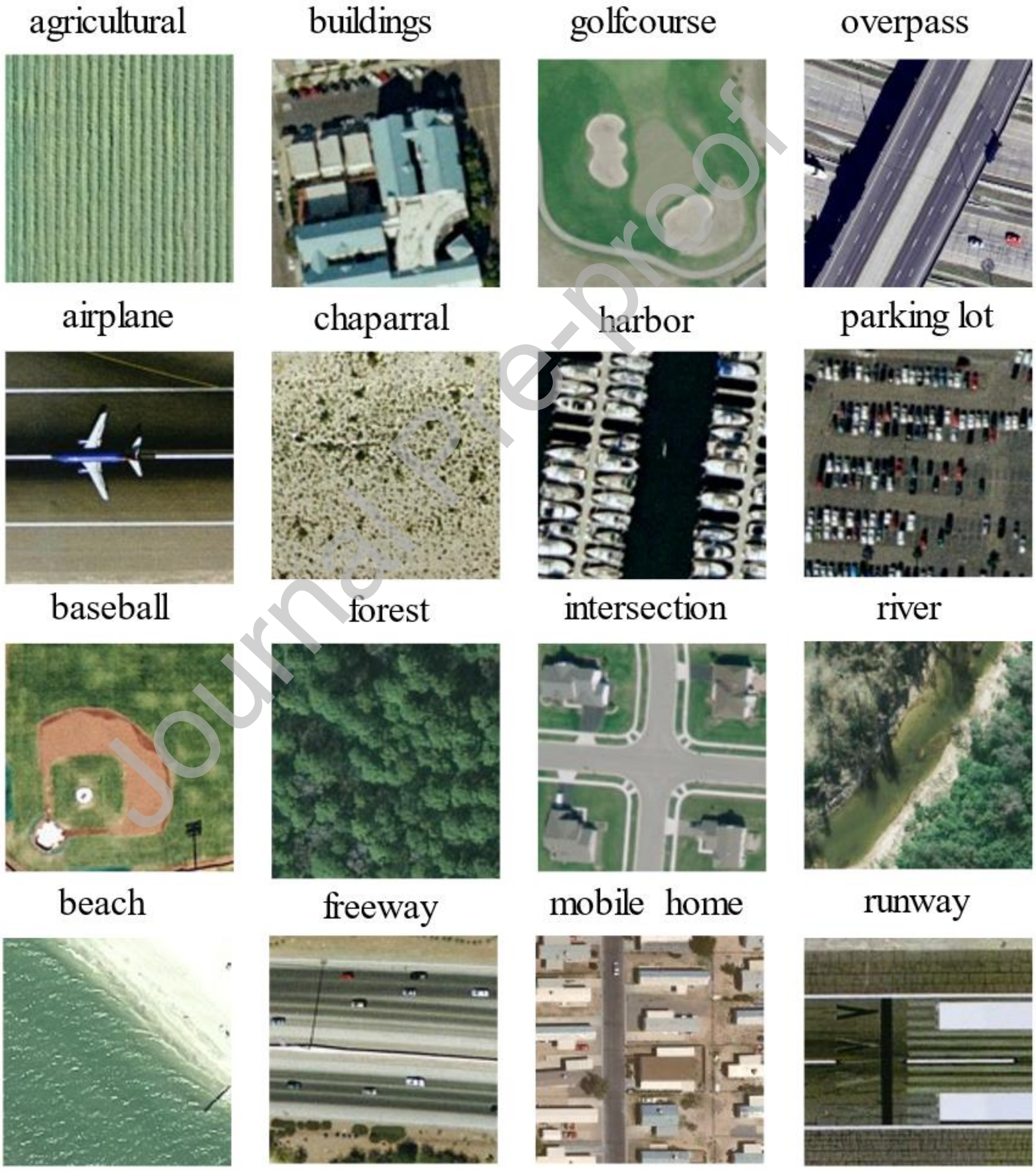


**Figure 6:** Sample images from the UCMerced_LandUse dataset.

### 4.2. Evaluation Parameters

Herein, we used metrics such as recall, precision, F1-score, and accuracy to assess model efficacy. These metrics are computed from the fundamental outcomes: False Negatives (FN), False Positives (FP), True Positives (TP), and True Negatives (TN). Accuracy depicts the relative amount of correctly classified occurrences to the total, recall measures the correct prediction of positive cases, precision quantifies the accuracy of positive predictions, and the F1-score is the reciprocal of the arithmetic mean of precision and recall. Following are the mathematical formulas of these metrics:

**Table 2.** Hyperparameters of different models.

| Performance measure | DenseNet-201 | EfficientNet-V27 | Xception | EfficientNet-B0 | Inception-V3 | ResNet-152 | MobileNet V2 |
|---|---|---|---|---|---|---|---|
| Image size | 224*224 | 224*224 | 224*224 | 224*224 | 224*224 | 224*224 | 224*224 |
| Optimizer | SGD | SGD | SGD | SGD | SGD | SGD | SGD |
| Batch Size | 32 | 32 | 32 | 32 | 32 | 32 | 32 |
| Epochs | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| Loss | CC | CC | CC | CC | CC | CC | CC |
| Activation Function | SoftMax | SoftMax | SoftMax | SoftMax | SoftMax | SoftMax | SoftMax |

$$\text{Precision} = \frac{TP}{TP+FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{8}$$

$$\text{F1} - \text{Score} = 2 \times \left(\frac{pre \times \text{rec}}{pre + \text{rec}}\right) \tag{9}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

### 4.3. Ablation study

We conducted an ablation study to evaluate the impact of different components and models on overall performance. This study is crucial for understanding which elements contribute most significantly to model accuracy and optimizing the model accordingly. Initially, we experimented with several deep-learning models individually to determine their solo performance. The models tested included ResNet152, MobileNetV2, DenseNet121, EfficientNetV2S, Xception, DenseNet201, and EfficientNetB0. As shown in **Table 4**, the accuracy achieved by each model on the UCMerced dataset revealed that DenseNet201, EfficientNetV2S, and Xception were the top performers, each achieving an accuracy of 95%. Building on these findings, we explored the potential of ensemble methods to further enhance accuracy. We first created two-model ensembles, such as EfficientNetV2S & Xception, Xception & DenseNet201, and DenseNet201 & EfficientNetV2S. Each of these combinations maintained a high accuracy of 95%, demonstrating the robustness of the individual models when combined. We then tested a three-model ensemble comprising EfficientNetV2S, Xception, and DenseNet201, which resulted in a slight improvement, achieving an accuracy of 96%. Finally, we incorporated an attention mechanism into the model, which further increased the overall accuracy to 97%. After applying soft attention, the ensemble model misclassified approximately 3% of land cover images primarily due to challenging environmental factors such as varying lighting conditions, terrain complexity, and seasonal changes. Lighting conditions can significantly affect the appearance of land cover types in remote sensing images, leading to confusion between classes, especially in areas with shadows or low sunlight. Additionally, terrain complexity, such as steep slopes or heterogeneous

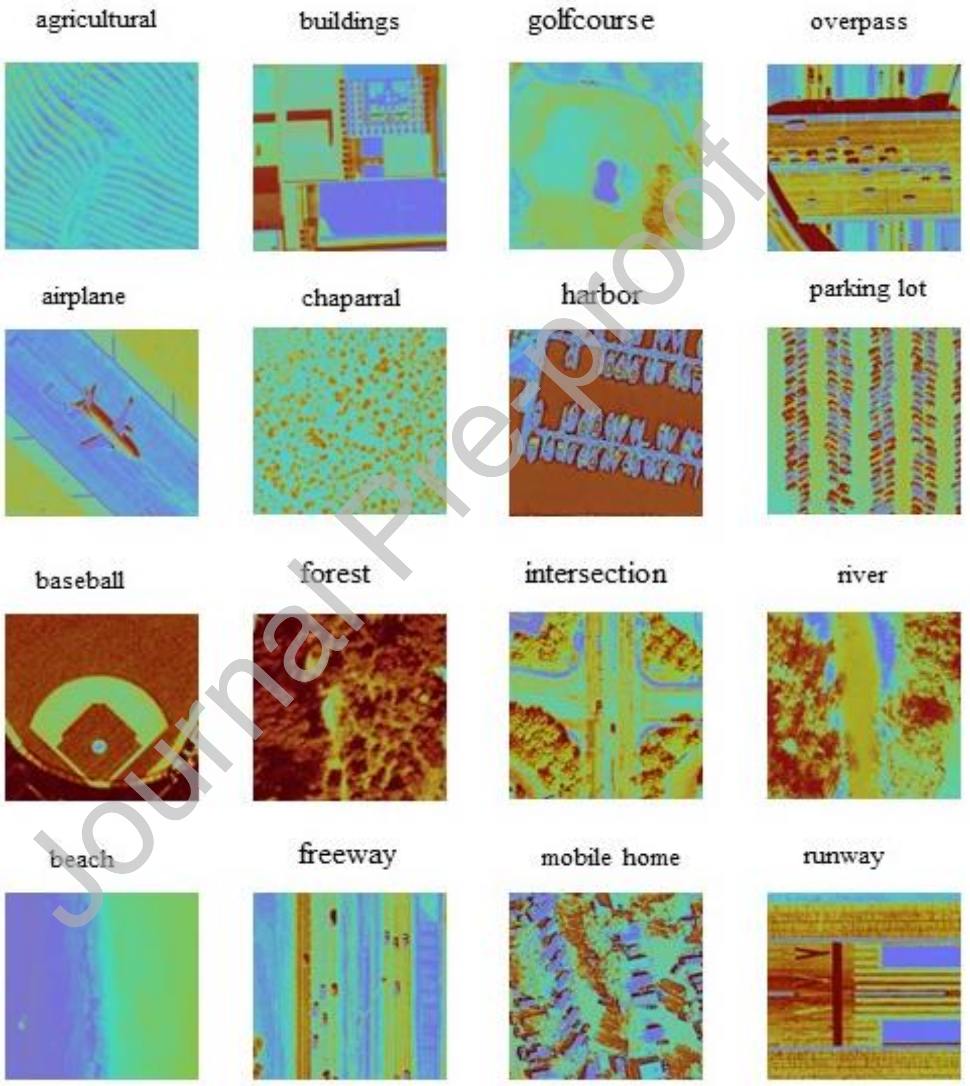


**Figure 7:** Grad Cam visualization of the proposed model for different classes

landscapes, can create mixed pixels where multiple land cover types are present, making accurate classification difficult. Seasonal variations, such as changes in vegetation cover or water levels, can also alter the spectral characteristics of certain land cover types, further contributing to misclassification.

These factors highlight the inherent difficulties in remote sensing image analysis and underscore the need for continued refinement of models to improve their robustness under diverse conditions. This progression highlights the benefits of combining high-performing models and leveraging attention mechanisms to boost performance, as demonstrated in **Table 6**. The proposed ensemble model's performance is also assessed using a confusion matrix.

**Table 3.** Statistical details of the UCMerced_LandUse dataset [54].

| Class | No of images | Training images | Testing images |
|---|---|---|---|
| agriculture | 1000 | 700 | 300 |
| airplane | 1000 | 700 | 300 |
| Baseball diamond | 1000 | 700 | 300 |
| Beach | 1000 | 700 | 300 |
| Building | 1000 | 700 | 300 |
| Chaparral | 1000 | 700 | 300 |
| Forest | 1000 | 700 | 300 |
| Freeway | 1000 | 700 | 300 |
| Golf course | 1000 | 700 | 300 |
| Harbor | 1000 | 700 | 300 |
| Intersection | 1000 | 700 | 300 |
| Mobile home park | 1000 | 700 | 300 |
| Overpass | 1000 | 700 | 300 |
| Parking lot | 1000 | 700 | 300 |
| River | 1000 | 700 | 300 |
| Runway | 1000 | 700 | 300 |
| Storage tanks | 1000 | 700 | 300 |
| Tennis court | 1000 | 700 | 300 |
| **Total images** | **18000** | **12600** | **5400** |

**Table 4.** Ablation study results of different models.

| Model | Attention | UCMerced Dataset |
|---|---|---|
| ResNet152 | × | 89 |
| MobileNetV2 | × | 88 |
| DenseNet121 | × | 88 |

| | | |
|---|---|---|
| EfficientNetV2S | × | 95 |
| Xception | × | 95 |
| DenseNet201 | × | 95 |
| EfficientNetB0 | × | 90 |
| EfficientNetV2S& Xception | × | 95 |
| Xception & DenseNet201 | × | 95 |
| DenseNet201 & EfficientNetV2S | × | 95 |
| EfficientNetV2S& Xception & DenseNet201 | × | 96 |
| Proposed (soft Attention) | ✓ | 97 |

To further evaluate the effectiveness of the individual models and the proposed ensemble method, we examined precision, recall, and F1-score metrics. These metrics provide a comprehensive view of the model performance beyond mere accuracy. **Table 5** shows the evaluation results for each model.

**Table 5.** Evaluating the suggested deep ensemble approach against contemporary refined techniques.

| Models | Precision% | Recall% | F1-Score% |
|---|---|---|---|
| Inception-V3 [48] | 95 | 96 | 95 |
| VGG16 [55] | 97 | 95 | 95 |
| NasNetLarge [56] | 58 | 58 | 57 |
| ResNet101 [57] | 96 | 95 | 96 |
| Xception [50] | 98 | 97 | 96 |
| MobileNetV2 [49] | 96 | 95 | 96 |
| DenseNet201 [58] | 94 | 93 | 93 |
| **Proposed** | **96** | **96** | **97** |

The results in **Table 5** indicate that while individual models like Xception and MobileNetV2 perform exceptionally well, the proposed ensemble model slightly surpasses them in F1-score, reaching 97%. This underscores the effectiveness of our ensemble strategy in leveraging the strengths of multiple models to achieve superior overall performance. **Figure 7** depicts the Grad-CAM visualizations of all classes, highlighting the activation regions that the models focus on during predictions. **Figure 8** presents the ROC curve representation for each class, illustrating the model's discriminative power.

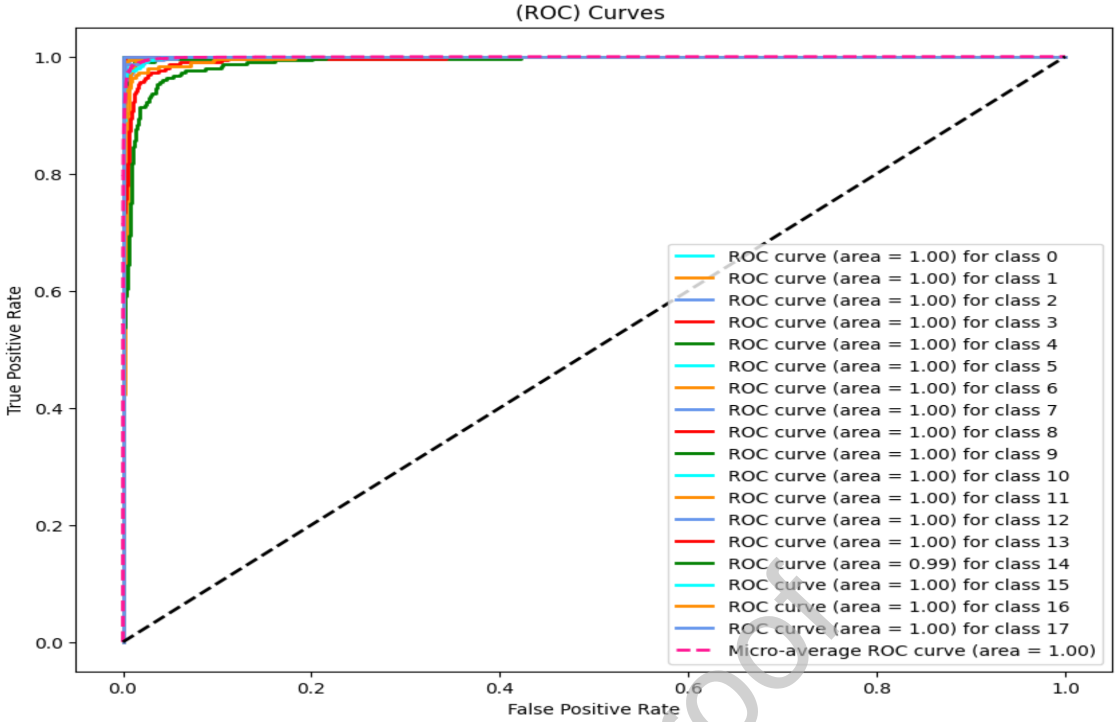
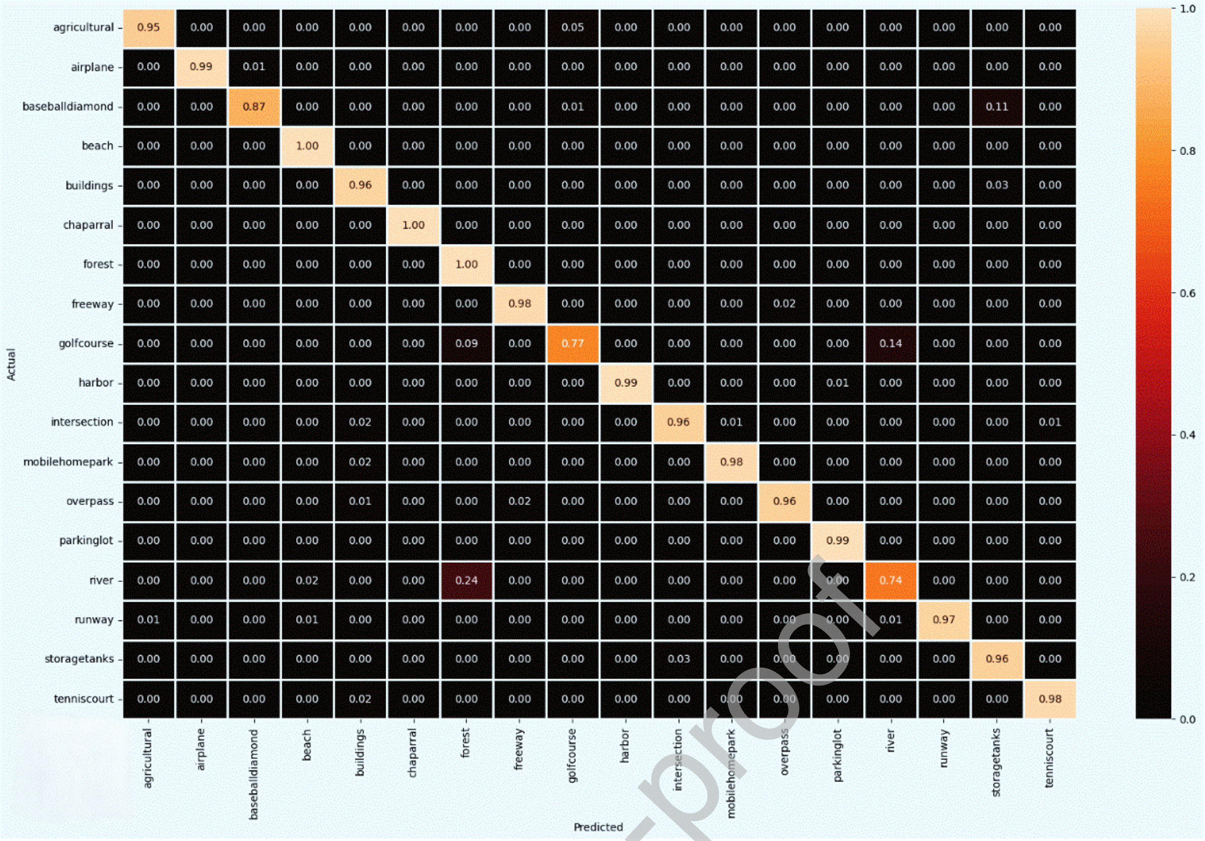


**Figure 8.** Proposed model ROC representation of each class and average.

ROC curves demonstrate a high area under the curve (AUC) values, which is consistent with the high accuracy observed in the confusion matrix. This further supports the robustness and effectiveness of the model in distinguishing between the different classes. Additionally, **Figure 9** shows the confusion matrix of the proposed model, providing insights into the classification accuracy. The confusion matrix provided highlights the model's strong classification performance, with high accuracy across most classes. The matrix clearly illustrates the model’s ability to correctly predict the majority of instances, with minimal misclassification. This supports the overall effectiveness and reliability of the model in handling the given classification task.



| | agricultural | airplane | baseballdiamond | beach | buildings | chaparral | forest | freeway | golfcourse | harbor | intersection | mobilehomepark | overpass | parkinglot | river | runway | storagetanks | tenniscourt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| agricultural | 0.95 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| airplane | 0.00 | 0.99 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| baseballdiamond | 0.00 | 0.00 | 0.87 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 |
| beach | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| buildings | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 |
| chaparral | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| forest | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| freeway | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.98 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| golfcourse | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 0.77 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| harbor | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| intersection | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| mobilehomepark | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.98 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| overpass | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| parkinglot | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 |
| river | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.24 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.74 | 0.00 | 0.00 | 0.00 |
| runway | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.97 | 0.00 | 0.00 |
| storagetanks | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.00 |
| tenniscourt | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.98 |

**Figure 9**: Depicts a confusion matrix of the proposed model.

### 4.4. Comparative Analysis

We conducted a comparative analysis to evaluate the performance of our proposed model against several baseline models from recent literature. **Table 6** presents the comparative analysis accuracy percentages of various ensemble models from different studies, along with the analysis of the strengths and limitations of each method in our proposed model. For instance, Ali Jamal et al. [59] achieved an accuracy of 87.50% using an ensemble of Inception, ResNet, and GoogleNet. This ensemble combines

**Table 6:** Comparative analysis of the proposed model with baselines.

| Model | Accuracy% | Advantages | Disadvantages |
|---|---|---|---|
| Ali Jamal et al [59] (Inception, ResNet, Google Net) | 87.50 | Combines complementary architectures for improved feature extraction | Lower accuracy, limited to specific datasets |
| Rahee Walambe et al [32] (RetinaResNet50, YOLOV3, SSD, ResNet) | 89.00 | Robust, effective for object detection | Computationally expensive due to complexity |

| | | | |
|---|---|---|---|
| Runyu Fan et al. [3] (ResNet50, ResneXt, ShuffleNet-V2) | 94.0 | Lightweight models, fast feature extraction | Struggles with more complex datasets |
| Sangdaow Noppitak et al [56] (NASNetLarge, VGG16, VGG19) | 92.80 | Strong classification skills with deep networks | High computational resources and training time required |
| Lei Song et al. [30] (CNN, ViT) | 93.68 | Transformer-based models excel in sequential image analysis | Requires large datasets for optimal performance |
| Jing chen et al. [31] (CMFM, MFM) | 93.64 | Effective for multi-modal fusion | Limited generalizability to other tasks |
| Kai Hu et al. [60] | 91.1 | -- | -- |
| Xu et al [47] (ATFM, DAL, FSGL) | 52.43 | | |
| Ma et al [17] (Hybird FENet) | 82.85 | Advanced hybrid models providing good accuracy in specific cases | May lacks generalization across different dataset types |
| Li et al [18] (Hybird EMFNet) | 96.1 | -- | -- |
| Pan et al. [19] (CNN) | 95.5 | | |
| **Proposed (**DenseNet201, EfficientNetV2S, Xception**)** | **97.0** | Excellent accuracy, efficient feature extraction, scalable to complex datasets | Higher processing resources required, extended training time |

complementary architecture for improved feature extraction; however, it shows lower accuracy and may be limited to specific datasets. Similarly, Rahee Walambe et al. [32] reported an accuracy of 89.00% with an ensemble comprising RetinaResNet50, YOLOV3, and SSD ResNet. While this model is robust and performs well for object detection, it is computationally expensive due to its complexity. Runyu Fan et al. [3] obtained a higher accuracy of 94% with a combination of ResNet50, ResneXt, and ShuffleNet-V2. This ensemble benefits from lightweight models and rapid feature extraction, but it may struggle with more complicated datasets. Sangdaow Noppitak et al. [56] used an ensemble of NASNetLarge, VGG16, and VGG19, achieving an accuracy of 92.80%. While these deep networks offer significant classification skills, they come at the cost of large computing resources and training time. In contrast, Lei Song et al. [30] achieved 93.68% accuracy using a combination of CNN and

ViT, benefiting from transformer-based models that excel in sequential image analysis but require large datasets for optimal performance. Jing Chen et al. [31] also achieved 93.64% accuracy using CMFM and MFM, which are effective for multi-modal fusion but may have limited generalizability to other tasks. Ma and Li et al. [17], [18]  achieved the accuracy of 82.85 and 96.1 by using advanced hybrid FENet and hybrid EMFNet models, respectively.  We acknowledge the need for a more detailed comparison of existing literature and have addressed this by adding a comprehensive comparison that highlights the performance indicators, strengths, and limitations of various ensemble models discussed in recent studies. Table 6 now includes a detailed analysis of each model's architecture, accuracy, and advantages and disadvantages. For instance, models such as Inception, ResNet, and GoogleNet perform well in feature extraction but are limited by dataset specificity. Similarly, models like RetinaResNet50 and YOLOV3 provide strong object detection capabilities but come at a high computational cost.

In contrast, our proposed ensemble model, which integrates DenseNet201, EfficientNetV2S, and Xception, with spatial attention module significantly outperformed these baselines, achieving an accuracy of 97%. This combination provides for excellent accuracy, efficient feature extraction, and scalability, making the model applicable to huge and complicated datasets. However, the integration of these complex architectures increases the processing resources and extends the training period. Nevertheless, the improved performance of our approach, particularly in applications requiring high accuracy and dependability in image categorization, highlights its efficacy.

### 4.5. Computational Complexity Analysis

In addition to evaluating the accuracy of the proposed ensemble model, it is crucial to assess its computational complexity compared to existing methods. **Table 7** provides a comparison of various models in terms of the number of parameters (in millions), model size (in MB), and latency (in seconds). This analysis helps to understand the trade-offs between accuracy and computational efficiency. ResNet-101 [57] has 44.7 million parameters and a model size of 171 MB, with a latency of 90 seconds. Xception [50] and Inception-V3 [48], [61] have similar computational complexities, with Xception having 22.9 million parameters, an 88 MB model size, and 90 seconds latency, while Inception-V3 has 23.9 million parameters, a 92 MB model size, and 67 seconds latency. MobileNet-V2 [49] is significantly lighter, with 3.5 million parameters and a 14 MB model size, but still has a latency of 87 seconds. DenseNet-201 [62], [58] and VGG16 [55] have higher computational complexities, with DenseNet-201 having 20.2 million parameters, an 80 MB model size, and 92 seconds latency, and VGG16 having 138.4 million parameters, a 528 MB model size, and 93 seconds latency. EfficientNet-B0 [51] has 5.3 million parameters and a 29 MB model size, and it has a latency of 84 seconds. Our proposed ensemble model, which combines DenseNet201, EfficientNetV2S, and

Xception, exhibits a total model size of 247 MB and a latency of 32 seconds with 64.4 million effective parameters, While the proposed model has a larger size compared to individual models like MobileNet-V2 [49] and EfficientNet-B0 [61], it significantly reduces latency, demonstrating an effective balance between computational complexity and performance. This balance is crucial for practical applications where both high accuracy and reasonable computational efficiency are required.

**Table 7:** Compares the proposed method's computational complexity with existing methods.

| Methods | Parameters (in million) | Model size (MB) | Latency(sec) |
|---|---|---|---|
| ResNet-101 [57] | 44.7 | 171 | 90 |
| Xception [50] | 22.9 | 88 | 90 |
| Inception-V3 [48] | 23.9 | 92 | 67 |
| MobileNet-V2 [49] | 3.5 | 14 | 87 |
| DenseNet-201 [62] | 20.2 | 80 | 92 |
| VGG16 [55] | 138.4 | 528 | 93 |
| EfficientNet-B0 [51] | 5.3 | 29 | 84 |
| **Proposed** | **64.7** | **247** | **32** |

The results presented in **Table 8** reflect the K-fold cross-validation performance of the proposed model and the second-best model from the ablation study. The accuracies reported are averaged across five different folds, ensuring robustness against variability from different parameter initializations. We also conducted paired t-tests to evaluate the statistical significance of the performance differences. The t-values and p-values demonstrate that the improvements of the proposed model over the second-best model are statistically significant ($p < 0.05$) across all folds, underscoring the reliability of our approach.

**Table 8:** K fold Cross Validation of proposed model and second-best model in Ablation study.

| Fold | Proposed model Acc | Second-best Acc | T_Value | P_value |
|---|---|---|---|---|
| Fold 1 | 96.8 | 95.8 | 15.0 | 0.042 |
| Fold 2 | 97.0 | 96.1 | 14.5 | 0.043 |
| Fold 3 | 96.9 | 95.8 | 15.0 | 0.042 |
| Fold 4 | 97.3 | 96.2 | 16.0 | 0.039 |
| Fold 5 | 97.1 | 96.2 | 15.5 | 0.041 |

| | | |
|---|---|---|
| **Mean Accuracy** | **97.02** | **96.00** |
| **Standard Deviation** | **0.17** | **0.18** |

To calculate the t-value and p-value for comparing two models, we perform a paired t-test. We start by collecting the accuracy results from both models across the same dataset folds. Next, we compute the difference between the paired accuracies for each fold. After that, we calculate the mean and standard deviation of these differences. Using these values, we then compute the t-value, which indicates the significance of the difference between the models' performances. Finally, we calculate the p-value, which helps us determine if the difference is statistically significant (e.g., $p < 0.05$). This process allows us to rigorously evaluate whether the proposed model significantly outperforms the baseline.

## 5. Conclusion and Future Direction

The research presented in this paper has highlighted substantial advancements in UAV image scene classification through the application of ensemble learning techniques. By combining DenseNet-201, EfficientNetV2S, and Xception models with a spatial attention module, we achieved an impressive validation accuracy of 97% on the UC Merced Land Use Dataset. This result underscores the effectiveness of integrating diverse deep-learning models and leveraging spatial attention to enhance classification performance and robustness significantly. Our ensemble model benefits from DenseNet-201's efficient feature reuse, EfficientNetV2S's scalable design, and Xception's depth-wise separable convolutions. The addition of the spatial attention module further refines the model's ability to focus on relevant regions within the images, thereby improving accuracy and interpretability. This comprehensive approach not only optimizes classification performance but also enhances the model's adaptability to complex and varied scenes.

In addition to the scientific contributions, the practical applications of this study are particularly promising in fields such as agricultural monitoring and environmental protection. In agricultural surveillance, UAVs coupled with our suggested model may effectively categorize and monitor wide landscapes, recognizing patterns in crop health, detecting pest infestations, and managing resources more efficiently. The great precision of the algorithm in discriminating across various situations offers quick and relevant information, which may help farmers improve yields and decrease waste. Similarly, in environmental protection, the model may be used for monitoring ecosystems, identifying deforestation, tracking land use changes, and measuring the health of forests and wetlands. The capacity to comprehend varied and complicated environmental scenarios with high accuracy makes it beneficial for conservation efforts and for responding to environmental changes quickly. These practical applications offer hope for a more efficient and sustainable future.

However, despite these promising results, there are limitations that should be addressed in future research to further improve the robustness and applicability of our model. While the ensemble model demonstrated high accuracy on the UC Merced Land Use Dataset, the relatively small size and limited diversity of this dataset may not fully capture the complexity of real-world scenes. Expanding the dataset to include a broader range of environments and more varied imagery will be crucial for improving the generalizability of the model. Additionally, the current approach focuses primarily on spatial information, with limited consideration of temporal dynamics. Integrating spatial-temporal context could significantly enhance performance in dynamic and changing environments, which is critical for real-time UAV applications. Furthermore, the computational complexity of the ensemble model presents a challenge, especially for deployment on resource-constrained edge devices. Future work will focus on optimizing the model for real-time deployment by fine-tuning parameters, reducing computational overhead, and exploring advanced transfer learning techniques to improve adaptability to various conditions. By addressing these limitations and expanding its practical use, our proposed approach holds great potential for revolutionizing UAV-based monitoring across industries, particularly in agriculture and environmental protection, ultimately leading to more efficient and sustainable management of natural resources.

## Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and by Basic Science Research Prograsm through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2024-00461244) and by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Digital Breeding Transformation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (322063-03-1-SB010).

## References

1. Bikkasani, R.H., M. Dhanya, and S. Veena. *Analysis of Long-term Changes for Land Use and Land Cover using Machine Learning: A case study*. in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. 2023. IEEE.
2. Shakhatreh, H., et al., *Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges.* Ieee Access, 2019. **7**: p. 48572-48634.
3. Yao, H., R. Qin, and X. Chen, *Unmanned Aerial Vehicle for Remote Sensing Applications—A Review.* Remote Sensing, 2019. **11**(12): p. 1443.
4. Duffy, J.P., et al., *Location, location, location: considerations when using lightweight drones in challenging environments.* Remote Sensing in Ecology and Conservation, 2018. **4**(1): p. 7-19.

5. Webb, G.I. and Z. Zheng, *Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques.* IEEE Transactions on Knowledge and Data Engineering, 2004. **16**(8): p. 980-991.
6. Joshi, G.P., et al., *Ensemble of deep learning-based multimodal remote sensing image classification model on unmanned aerial vehicle networks.* Mathematics, 2021. **9**(22): p. 2984.
7. Fayaz, M., et al., *Land-Cover Classification Using Deep Learning with High-Resolution Remote-Sensing Imagery.* Applied Sciences, 2024. **14**(5): p. 1844.
8. Temenos, A., et al., *Interpretable deep learning framework for land use and land cover classification in remote sensing using SHAP.* IEEE Geoscience and Remote Sensing Letters, 2023. **20**: p. 1-5.
9. Jockusch, O., et al. *Generative AI-based Land Cover Classification via Federated Learning CNNs: Sustainable Insights from UAV Imagery*. in *2024 IEEE Conference on Technologies for Sustainability (SusTech)*. 2024. IEEE.
10. Sefercik, U.G., et al., *3D positioning accuracy and land cover classification performance of multispectral RTK UAVs.* International Journal of Engineering and Geosciences, 2023. **8**(2): p. 119-128.
11. Fu, H., et al., *Three-dimensional singular spectrum analysis for precise land cover classification from UAV-borne hyperspectral benchmark datasets.* ISPRS Journal of Photogrammetry and Remote Sensing, 2023. **203**: p. 115-134.
12. Mollick, T., M.G. Azam, and S. Karim, *Geospatial-based machine learning techniques for land use and land cover mapping using a high-resolution unmanned aerial vehicle image.* Remote Sensing Applications: Society and Environment, 2023. **29**: p. 100859.
13. Liu, Y., et al., *Counterfactual-augmented few-shot contrastive learning for machinery intelligent fault diagnosis with limited samples.* Mechanical Systems and Signal Processing, 2024. **216**: p. 111507.
14. Papoutsis, I., et al., *Benchmarking and scaling of deep learning models for land cover image classification.* ISPRS Journal of Photogrammetry and Remote Sensing, 2023. **195**: p. 250-268.
15. Jin, B. and X. Xu, *Contemporaneous causality among price indices of ten major steel products.* Ironmaking & Steelmaking, 2024: p. 03019233241249361.
16. Jin, B. and X. Xu, *Machine learning predictions of regional steel price indices for east China.* Ironmaking & Steelmaking, 2024: p. 03019233241254891.
17. Ma, Z., et al., *FENet: Feature enhancement network for land cover classification.* International Journal of Remote Sensing, 2023. **44**(5): p. 1702-1725.
18. Li, C., R. Hang, and B. Rasti, *EMFNet: Enhanced multisource fusion network for land cover classification.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021. **14**: p. 4381-4389.
19. Pan, S., et al., *Land-cover classification of multispectral LiDAR data using CNN with optimized hyper-parameters.* ISPRS Journal of Photogrammetry and Remote Sensing, 2020. **166**: p. 241-254.
20. Kwan, C., et al., *Deep learning for land cover classification using only a few bands.* Remote Sensing, 2020. **12**(12): p. 2000.
21. Zhang, C., et al., *A multi-level context-guided classification method with object-based convolutional neural network for land cover classification using very high resolution remote sensing images.* International Journal of Applied Earth Observation and Geoinformation, 2020. **88**: p. 102086.
22. Rajendran, G.B., et al., *Land-use and land-cover classification using a human group-based particle swarm optimization algorithm with an LSTM Classifier on hybrid pre-processing remote-sensing images.* Remote Sensing, 2020. **12**(24): p. 4135.

23. Chatterjee, A., et al., *Unsupervised land cover classification of hybrid and dual-polarized images using deep convolutional neural network.* IEEE Geoscience and Remote Sensing Letters, 2020. **18**(6): p. 969-973.
24. MOON, G.-S., K.-S. KIM, and Y.-J. CHOUNG, *Land cover classification based on high resolution KOMPSAT-3 satellite imagery using deep neural network model.* Journal of the Korean Association of Geographic Information Studies, 2020. **23**(3): p. 252-262.
25. Jin, B. and X. Xu, *Carbon emission allowance price forecasting for China Guangdong carbon emission exchange via the neural network.* Global Finance Review, 2024. **6**(1): p. 3491-3491.
26. Xu, X. and Y. Zhang, *Corn cash price forecasting with neural networks.* Computers and Electronics in Agriculture, 2021. **184**: p. 106120.
27. Jin, B. and X. Xu, *Forecasting wholesale prices of yellow corn through the Gaussian process regression.* Neural Computing and Applications, 2024. **36**(15): p. 8693-8710.
28. Jin, B. and X. Xu, *Pre-owned housing price index forecasts using Gaussian process regressions.* Journal of Modelling in Management, 2024.
29. Aspri, M., G. Tsagkatakis, and P. Tsakalides, *Distributed training and inference of deep learning models for multi-modal land cover classification.* Remote Sensing, 2020. **12**(17): p. 2670.
30. Song, L., et al., *Axial cross attention meets CNN: Bibranch fusion network for change detection.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022. **16**: p. 21-32.
31. Chen, J., et al., *Double branch parallel network for segmentation of buildings and waters in remote sensing images.* Remote Sensing, 2023. **15**(6): p. 1536.
32. Walambe, R., A. Marathe, and K. Kotecha, *Multiscale Object Detection from Drone Imagery Using Ensemble Transfer Learning. Drones 2021, 5, 66*. 2021, s Note: MDPI stays neu-tral with regard to jurisdictional claims in ….
33. Subedi, M.R., et al., *Ensemble Machine Learning on the Fusion of Sentinel Time Series Imagery with High-Resolution Orthoimagery for Improved Land Use/Land Cover Mapping.* Remote Sensing, 2024. **16**(15): p. 2778.
34. Jin, B. and X. Xu, *Wholesale price forecasts of green grams using the neural network.* Asian Journal of Economics and Banking, 2024.
35. Jin, B. and X. Xu, *Price forecasting through neural networks for crude oil, heating oil, and natural gas.* Measurement: Energy, 2024. **1**(1): p. 100001.
36. Jin, B. and X. Xu, *Forecasts of thermal coal prices through Gaussian process regressions.* Ironmaking & Steelmaking, 2024: p. 03019233241265194.
37. Jin, B. and X. Xu, *Palladium Price Predictions via Machine Learning.* Materials Circular Economy, 2024. **6**(1): p. 32.
38. Fu, B., et al., *Comparison of RFE-DL and stacking ensemble learning algorithms for classifying mangrove species on UAV multispectral images.* International Journal of Applied Earth Observation and Geoinformation, 2022. **112**: p. 102890.
39. Song, L., A.B. Estes, and L.D. Estes, *A super-ensemble approach to map land cover types with high resolution over data-sparse African savanna landscapes.* International Journal of Applied Earth Observation and Geoinformation, 2023. **116**: p. 103152.
40. Colkesen, I. and M.Y. Ozturk, *A comparative evaluation of state-of-the-art ensemble learning algorithms for land cover classification using WorldView-2, Sentinel-2 and ROSIS imagery.* Arabian Journal of Geosciences, 2022. **15**(10): p. 942.
41. Amin, S.U., et al., *An automated chest X-ray analysis for COVID-19, tuberculosis, and pneumonia employing ensemble learning approach.* Biomedical Signal Processing and Control, 2024. **87**: p. 105408.
42. McCoy, J., et al., *Ensemble Deep learning for sustainable multimodal UAV classification.* IEEE Transactions on Intelligent Transportation Systems, 2022.

43. Namoun, A., et al., *An ensemble learning based classification approach for the prediction of household solid waste generation.* Sensors, 2022. **22**(9): p. 3506.
44. Sefrin, O., F.M. Riese, and S. Keller, *Deep learning for land cover change detection.* Remote Sensing, 2020. **13**(1): p. 78.
45. Deepan, P. and L. Sudha. *Scene classification of remotely sensed images using ensembled machine learning models*. in *Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication: Proceedings of MDCWC 2020*. 2021. Springer.
46. Wambugu, N., et al., *A hybrid deep convolutional neural network for accurate land cover classification.* International Journal of Applied Earth Observation and Geoinformation, 2021. **103**: p. 102515.
47. Xu, R., et al., *RSSFormer: Foreground saliency enhancement for remote sensing land-cover segmentation.* IEEE Transactions on Image Processing, 2023. **32**: p. 1052-1064.
48. Szegedy, C., et al. *Rethinking the inception architecture for computer vision*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
49. Sandler, M., et al. *Mobilenetv2: Inverted residuals and linear bottlenecks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
50. Chollet, F. *Xception: Deep learning with depthwise separable convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
51. Tan, M. and Q. Le. *Efficientnet: Rethinking model scaling for convolutional neural networks*. in *International conference on machine learning*. 2019. PMLR.
52. Singh, D., V. Kumar, and M. Kaur, *Densely connected convolutional networks-based COVID-19 screening model.* Applied Intelligence, 2021. **51**: p. 3044-3051.
53. Ul Amin, S., et al., *An Efficient Attention-Based Strategy for Anomaly Detection in Surveillance Video.* Computer Systems Science & Engineering, 2023. **46**(3).
54. Russakovsky, O., et al., *Imagenet large scale visual recognition challenge.* International journal of computer vision, 2015. **115**: p. 211-252.
55. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
56. Noppitak, S. and O. Surinta, *Ensemble convolutional neural network architectures for land use classification in economic crops aerial images.* ICIC Express Letters, 2021. **15**(6): p. 531-543.
57. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
58. Huang, G., et al. *Densely connected convolutional networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
59. Jamali, A., et al., *Comparing solo versus ensemble convolutional neural networks for wetland classification using multi-spectral satellite imagery.* Remote Sensing, 2021. **13**(11): p. 2046.
60. Hu, K., et al., *MCSGNet: A encoder–decoder architecture network for land cover classification.* Remote Sensing, 2023. **15**(11): p. 2810.
61. Hussain, A., et al., *An Efficient and Robust Hand Gesture Recognition System of Sign Language Employing Finetuned Inception-V3 and Efficientnet-B0 Network.* Computer Systems Science & Engineering, 2023. **46**(3).
62. Shafiq, M. and Z. Gu, *Deep residual learning for image recognition: A survey.* Applied Sciences, 2022. **12**(18): p. 8972.

**Declaration of interests**

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: