Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 3 (2026) 843 - 854 e-ISSN: 2580-0760

Improving Remote Sensing Classification with Ensemble Learning and XAI-based Interpretability

Radius Tanone[1], Li-Hua Li[2*], Ramli Ahmad[3], Alok Kumar Sharma[4], Agus Cahyo Nugroho[5]

[1,2,3,5]*Department of Information Management, Chaoyang University of Technology, Taichung City, Taiwan*
[4]*Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung City, Taiwan*
[1]*Department of Informatics, Satya Wacana Christian University, Salatiga, Indonesia*
[3]*Fakultas Teknik Komputer,Universitas Hamzanwadi, Lombok Timur, Indonesia*
[5]*Department of Information System, Unika Soegijapranata, Semarang, Indonesia*

[*]*lhli@gm.cyut.edu.tw*

Abstract

The challenge of classifying remote sensing images primarily stems from the unclear image quality produced by satellites during data collection. Relying on multiple models to recognize remote sensing images can sometimes lead to suboptimal performance. To address this issue, this study integrates transfer learning and ensemble learning to enhance the accuracy of single-model classification. This research employed pre-trained models, including EfficientNetB7, Vision Transformer, and ConvNeXt, and evaluated them on benchmark datasets RSI-CB256 and NWPU RESISC45. The results demonstrate that ensemble learning significantly boosts model performance beyond that of individual models. For the RSI-CB256 dataset, the average and geometric mean ensemble methods achieved the highest accuracy of 0.9985. For another dataset, namely NWPU RESISC45, the geometric mean ensemble achieved the best accuracy of 0.9720. Furthermore, this study explores model transparency and interpretability through eXplainable Artificial Intelligence (XAI) techniques. In addition to transparency, this study uses Grad-CAM to identify critical regions influencing the model's predictions in remote sensing image classification tasks.

Keywords: ensemble learning; explainable artificial intelligence; Grad-CAM; remote-sensing; transfer learning

How to Cite: R. Tanone, L.-H. Li, R. Ahmad, A. Kumar Sharma, and A. Cahyo Nugroho, "Improving Remote Sensing Classification with Ensemble Learning and XAI-based Interpretability", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 3, pp. 843 - 854, Jun. 2026. *Permalink/*DOI: https://doi.org/10.29207/resti.v10i3.7121

Received: August 25, 2025
Accepted: May 23, 2026
Available Online: June 28, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Deep learning architectures like Convolutional Neural Networks (CNNs) [1], [2] and Transformers [3], being foundational to many models, make them de facto. Recent improvements have been observed in various fields, with remarkable achievements in subfields such as computer vision [4]. Deep learning techniques have demonstrated high effectiveness in various industries, including agriculture [5], [6], healthcare [7], and remote sensing. Deep learning models can be used for study in remote sensing image analysis. Despite these advances, remote sensing image classification [8], [9] using deep learning methods remains challenging due to the high variability and complexity of images produced by satellites. Many remote sensing images have low resolution, noise, or quality inconsistencies, which can negatively impact model performance. Consequently, trained models may fail to achieve optimal accuracy due to factors such as architectural design, parameter count, computational complexity, and hyperparameter tuning.

To address these challenges, many studies have increasingly employed models from scratch, pre-trained models, ensemble learning techniques [10], and customized architectures. Building a model from scratch is often resource-intensive and time-consuming; in contrast, transfer learning with pre-trained models offers a more efficient alternative. However, adapting pre-trained models to specific datasets often requires architectural or parameter modifications to achieve better generalization. Despite their efficiency, single-model approaches frequently yield sub-optimal results. Ensemble learning, on the other hand, seeks to combine the strengths of multiple models to improve robustness compared to any single model.

Regarding the use of deep learning in remote sensing, model transparency is an important issue. The interpretability of model predictions is essential, as subject matter experts rely on these outputs to make critical decisions. Ensuring trust in model outcomes requires methods that can explain how predictions are generated. XAI [11], [12] addresses this challenge by providing interpretability tools that reveal how models make decisions. Techniques called LIME [13] and SHAP [14], [15] have been widely used to enhance transparency. In the context of image-based tasks, Gradient-weighted Class Activation Mapping (Grad-CAM) [16] has emerged as a particularly effective XAI approach for visualizing salient image regions contributing to model predictions. For remote sensing specialists, such interpretability aids in validating model reliability and improving decision-making confidence.

Extensive research has been conducted on image classification, including remote sensing applications. Due to the limited availability of new satellite datasets, researchers often utilize open-access datasets to test and validate novel classification methods. Bharti et al. [17] employed pre-trained models combined with data augmentation and diverse loss functions to improve classification performance. Similarly, Pham et al. [18] explored transfer learning and attention-based neural networks for remote sensing classification. Transformer-based models have recently gained popularity for such tasks. For instance, Jannat and Willis [19] applied Swin Transformers to remote sensing datasets and achieved superior performance compared to conventional CNN architectures. Xia et al. [20] further introduced a Multi-Scale Vision Transformer (MS-ViT) that integrates multiple spatial information streams and hybrid attention mechanisms to enhance scene classification accuracy.

Moreover, Adegun et al. [21] conducted a comprehensive evaluation of CNN- and Transformer-based architectures for remote sensing image classification, revealing that deep learning models effectively capture diverse visual features in satellite imagery. Maracani et al. [22] proposed an ensemble learning framework that combines CNN, Transformer, and ConvNeXt models, demonstrating improved accuracy for plankton image classification. These findings underscore the potential of pre-trained and hybrid deep learning models in improving classification performance through transfer and ensemble learning.

Realizing the potential of pre-trained models to improve classification performance, this study adopts an ensemble learning approach that integrates multiple deep learning architectures via transfer learning. To achieve this goal, experiments were conducted using publicly available remote sensing datasets and several pre-trained models. The proposed methodology combines CNN, Transformer, and ConvNeXt models to enhance classification accuracy. Furthermore, this research incorporates XAI techniques, using Grad-CAM, to visualize the model's transparency. In accordance with the findings of Kakogeorgiou and Karantzalos [23], Grad-CAM demonstrates superior computational efficiency compared to LIME and Occlusion methods. Building on these insights, this study employs Grad-CAM to achieve model transparency and improve interpretability in remote sensing image classification tasks.

XAI encompasses artificial intelligence systems designed to ensure that human users maintain intellectual oversight of algorithmic decisions and the underlying techniques [24]. The primary objective of XAI is to enhance the interpretability and transparency of AI models by providing understandable justifications for predictions [25], [26]. In this research, ensemble learning is applied for remote sensing image classification and emphasizes the importance of model interpretability. Although pre-trained models with transfer learning have shown strong performance, their decision-making mechanisms remain complex and opaque. Therefore, improving the interpretability of ensemble models is crucial to ensure that users can comprehend and trust the classification results [7], [12]. To meet these requirements, the interpretation process in ensemble learning must be conducted rigorously to ensure both transparency and reliability.

Transfer learning [27], [28] is a powerful approach that leverages pre-trained models by retraining them for new tasks. In this study, several high-performing architectures were utilized, including EfficientNetB7, ViT-B16, ViT-L16, and ConvNeXt, all fine-tuned using transfer learning. Each model was expected to deliver strong baseline performance, after which ensemble learning was applied to improve classification results further.

Ensemble learning [29], [30] represents a practical strategy that combines predictions from multiple models to reduce prediction variance and enhance generalization. By aggregating architecture models, ensemble methods mitigate the weaknesses of individual architecture and achieve improved overall performance. Ensemble techniques can be categorized by the diversity of their base components, whether in the training data, the model design, or the prediction aggregation.

Previous studies have primarily relied on conventional transfer learning approaches using pretrained deep learning models. While these methods have demonstrated promising results on large-scale datasets, they often encounter challenges when applied to relatively small remote sensing images. These images exhibit limited feature diversity and pose difficulties in generalization. Prior works have focused on optimizing a single model or backbone architecture without considering model diversity to enhance robustness. Consequently, model overfitting and limited adaptability to variations in small remote sensing datasets remain significant issues.

To overcome these limitations, this study employs an ensemble-based approach to improve the performance of a final remote sensing classification prediction. Unlike previous research that relied solely on individual pretrained models, this work trains multiple base models on the same dataset with diverse feature configurations and hyperparameters. This diversity enables the models to capture complementary patterns in the data, effectively reducing overfitting and improving generalization. Moreover, several ensemble aggregation strategies, including simple averaging, weighted averaging, and geometric mean, are employed to integrate model predictions seamlessly. By adopting this approach, the study addresses the existing gap in demonstrating how ensemble learning can significantly enhance transfer learning performance, particularly for small and complex remote sensing images.

The main objectives of this study are as follows. First, this research uses multiple pre-trained deep learning models, combined via transfer learning, for remote sensing image classification. Furthermore, Gradient Centralization (GC) [31] was implemented to improve model generalization and training precision. Second, this research uses several ensemble techniques, namely average, weighted-average, and geometric-mean ensembles. These techniques were explored to determine the most effective aggregation strategy. Third, to advance model transparency and interpretability, this research employed Grad-CAM, which transparently visualizes the critical image regions that influence classification decisions. These visualizations reveal how the model generates predictions, allowing domain experts to assess model reasoning and validate classification outcomes.

2. Methods

In this study, a comprehensive research workflow was developed to systematically complete the investigation. The ensemble learning framework, which integrates CNN and Transformer-based models, is illustrated in Figure 1. This figure shows the overall flow of the proposed method, demonstrating how ensemble learning is combined with XAI to generate interpretable classification results.

The research workflow begins with the utilization of two benchmark datasets, RSI-CB256 and NWPU RESISC45. The next step involves image preprocessing. During preprocessing, the data are divided into training, testing, and validation sets with proportions of 60%, 24%, and 16%, respectively.

The RSI-CB256 dataset, EfficientNetB7, and ViT-B16 models are employed with an input image size of 256 × 256 pixels. For the NWPU RESISC45 dataset, ViT-L16 and ConvNeXt-Large models are utilized, with an image input size of 384 × 384 pixels. The transfer learning process also incorporates a tuning optimizer enhanced with GC.

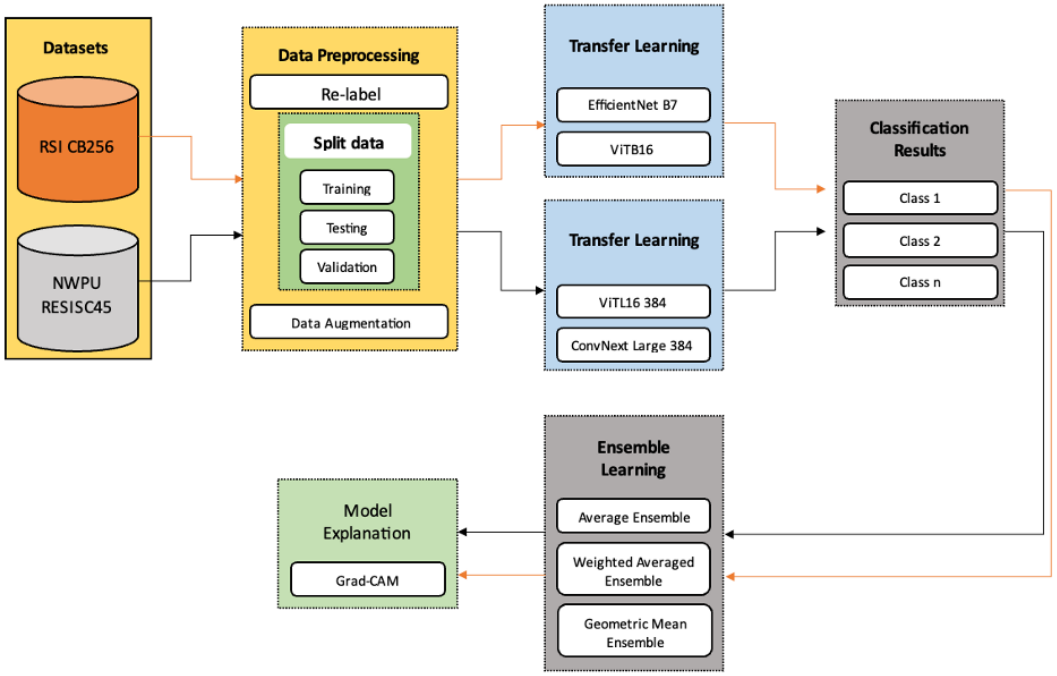


Figure 1. Research workflow

After obtaining the classification performance of each transfer learning model, ensemble learning is performed using three strategies namely average, weighted average, and geometric mean. The results of these approaches are then compared to identify the most effective ensemble method for achieving optimal classification accuracy.

Finally, model interpretability is achieved through Grad-CAM, which provides visual explanations of the model's classification behavior. This interpretability step helps experts understand model decisions and enhances confidence in the final classification outcomes.

2.1 Transfer Learning

Transfer learning [32] leverages knowledge acquired by a pretrained model on a large, well-labeled dataset to solve a related problem with limited data availability. Instead of training a model from scratch, this study applies transfer learning using state-of-the-art architectures such as EfficientNet, ConvNeXt, and Vision Transformer (ViT). EfficientNet and ConvNeXt are grounded in CNN, which excel at capturing local spatial dependencies and texture patterns. Contrarily, the ViT is designed to model long-range dependencies and global contextual information through self-attention mechanisms. By integrating these complementary architectures, the research leverages both the local feature-extraction strength of CNNs and the global representation capability of Transformers, yielding richer, more discriminative feature embeddings.

Building on these transfer learning foundations, ensemble learning is employed to further enhance classification robustness and accuracy. By combining multiple models trained under varying architectural configurations and hyperparameters, the ensemble approach mitigates individual model biases and captures diverse perspectives of the data. This synergy between transfer learning and ensemble learning enables the development of a more stable and high-performing classification framework, particularly well-suited for handling the complexity and variability inherent in remote sensing imagery.

EfficientNet [33] is a convolutional neural network architecture and scaling technique that employs a compound coefficient to uniformly scale all depth, width, and resolution dimensions of the image. For more information, refer to Figure 2.

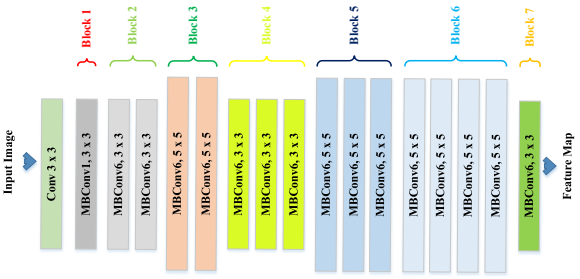


Figure 2. EfficientNet Arhitecture

ConvNeXT [34], a convolutional model inspired by ViT, aims to surpass their performance. As shown in Fig. 3, ConvNeXT retains the architectural simplicity of CNN while outperforming ViT in terms of accuracy, performance, and scalability. The CNN used in this study is based on the conventional ResNet-50 architecture, but certain modifications have been made to mimic the design of a ViT.

Vision Transformer (ViT) [35] employs a sequence-based representation for images, allowing it to predict class labels for the images. This approach enables the models to independently learn the visual structure of the images. The input images are processed patch-wise, where each patch is converted into a unified vector by aggregating the channels of all its pixels. This vector is then projected linearly to align with the specified input dimension. For reference, Figure 4 illustrates the ViT architecture.

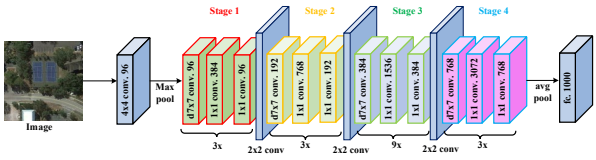


Figure 3. ConvNext Architecture

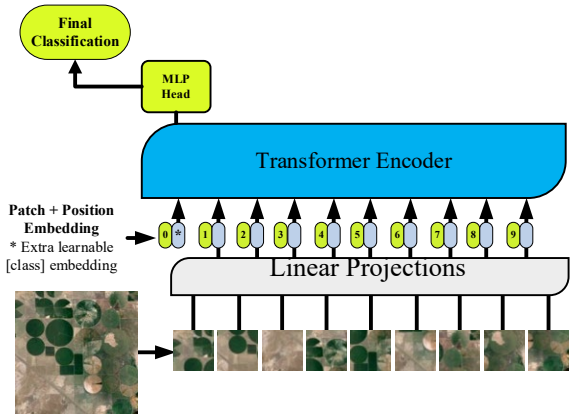


Figure 4. Vision Transformer Architecture

2.2 RMSProp Optimizer with GC

The gradient vectors are centrally located using the GC [31] technique to achieve a zero mean. The generalized GC method is an optimization technique that employs a constrained loss function and a projected gradient descent approach. In this work, this research employs transfer learning to train a deep learning model using RMSProp as the learning optimizer. The RMSProp expression is presented in Equations 1 and 2.

$$E[g^2]_{(t)} = \beta E[g^2]_{t-1} + (1-\beta)\left(\frac{\partial C}{\partial W}\right)^2 \quad (1)$$

$$W_{ij}(t) = W_{ij}(t-1) - \frac{\eta}{\sqrt{E[g^2]}}\frac{\partial C}{\partial W_{ij}} \quad (2)$$

$E[g^2]$ refers to the calculation of the moving average of squared gradients, the symbol $\frac{\partial C}{\partial W}$ represents the gradient of the cost function with respect to the weight, η refers to the learning rate, while the symbol β represents the moving average with a default value of 0.9. Then this research embedded GC into RMSProp. The math formula of GC can be expressed in Equation 3.

$$P^2 = P = P^T, \quad e^T P \nabla wL = 0 \quad (3)$$

The equations demonstrate that matrix P represents the projection of the hyperplane with the normal vector e onto the weight space. Additionally, $P\nabla wL$ denotes the gradient that has been projected. In addition, Equation 4 also allows for matrix formulation.

$$\Phi_{GC}(\nabla wL) = P\nabla wL, \qquad P = I - ee^T \tag{4}$$

In the practical implementation, it's possible to achieve global centering by directly subtracting the mean value from each weight vector.

2.3 Ensemble Learning Process

Averaging compilation combines the predictions of multiple base models into a single ensemble method [30]. Once the base models are trained, their predictions are combined into a single prediction using an averaging technique, such as geometric averaging, weighted averaging, or simple averaging. However, simple averaging doesn't consider the importance or effectiveness of each base model individually. Instead, it simply calculates the mean of its predictions.

Given a data set x=$\{x_1, \ldots, x_n\}$, the arithmetic mean (also mean or simple average), denoted $\bar{x}$, is the mean of the n values $x_1, \ldots, x_n$. Simple Averaging is defined by Equation 5.

$$\boldsymbol{x} = \frac{1}{n}\left(\sum_{i=1}^{n} \boldsymbol{x_i}\right) = \frac{x_1+x_2+\ldots+x_n}{n} \tag{5}$$

Weighted averaging [36], on the other hand, assigns each base model prediction a weight based on various factors and their performance in a validation set. To learn the weights, techniques such as cross-validation and grid search are usually used. Finding the geometric mean of all base model predictions using geometric mean averaging can be helpful when dealing with positive data, extreme values, or outliers.

The definition of the weighted average is as follows: with a non-empty finite tuple of data $(x_1, \ldots, x_n)$, and corresponding non-negative weights $(w_1, \ldots, w_n)$,, the weighted average is expressed in Equation 6.

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{6}$$

which expands to Equation (7).

$$\bar{\mathbf{x}} = \frac{w_1x_1+w_2x_2+\ldots+w_nx_n}{w_1+w_2+\ldots+w_n} \tag{7}$$

Note that the weights cannot be negative. However, some weights may be zero, but not all since this would result in division by zero. When the weights are normalized, see Equation 8.

$$\boldsymbol{w'_i} = \frac{w_i}{\sum_{j=1}^{n} w_j} \text{ and } \sum_{i=1}^{n} \boldsymbol{w'} = \mathbf{1} \tag{8}$$

then the weighted average can be expressed as a linear combination of the normalized weights and samples can be seen in Equation 9:

$$x' = \sum_{i=1}^{n} w'_i x_i \tag{9}$$

The concept of the geometric mean [37] pertains to a measure of central tendency for a finite collection of positive real numbers. Unlike the arithmetic mean that relies on their sum, the geometric mean is determined by calculating the product of these values (see Equation 10).

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \ldots x_n} \tag{10}$$

This research can also express the geometrics mean as the log scale arithmetic mean in Equation 11.

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = exp\left(\frac{1}{n}\sum_{i=1}^{n} 1_n a_i\right) \tag{11}$$

In this approach, the geometric mean of each base model's predictions is calculated. To reduce the risk of overfitting, individual base models are trained using different hyperparameters, methods, or feature subsets within the same dataset. This allows for the capture of various aspects of the data. After training, the predictions from the base models are consolidated into a unified prediction using the geometric mean.

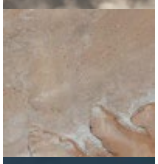
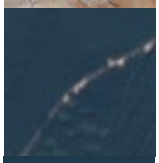
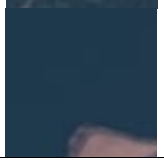
2.4 Dataset and Environment Setting

This experiment employed a benchmark dataset of remote sensing images sourced from Kaggle and TensorFlow. The dataset consisted of the open-source RSI-CB256 [38] and NWPU RESISC45 [39] datasets. For further details about each dataset, refer to Table 1-3.

Table 1. Dataset Details

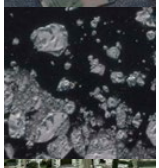
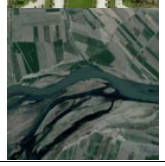
Item	RSI-CB256	NWPU RESISC45
Number of Class	4	45
Number of Image	5631	31500
Size	256 x 256	256 x 256

Table 2. Example samples from the RSI-CB256 dataset

Label	Image
Cloud	
Desert	
Forest	
Sea Lake	

In addition, the experimental setup provides detailed specifications for executing the training model. Specifically, the 11th Generation Intel(R) Core (TM) i9-11900K processor, operating at a frequency of 3.50 GHz and equipped with 16 CPUs, has been selected. The allocated memory is 32768 MB, while the graphics processing unit (GPU) is a NVIDIA GeForce RTX 3080 model with a total memory capacity of 26399 MB.

Table 3. Example samples from the NWPU RESISC45 dataset

Label	Image
Church	
Sea Ice	
Palace	
River	

2.5 Metrics Evaluation

Following the use of transfer learning for classification purposes on remote sensing images, the performance of the model is assessed by various assessment metrics, including precision, recall, accuracy, and F1 scores [40], [41]. The mathematical representation for these metrics is presented as follows (Equations 12-15):

$$Prec = \frac{TP}{TP+FP} \quad (12)$$

$$Rec = \frac{TP}{TP+FN} \quad (13)$$

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

$$F1 = 2\,x\,\frac{Rec\ x\ Prec}{Rec+Prec} \quad (15)$$

To evaluate the balance between classification performance and computational efficiency, a Trade-Off Score (TOS) is defined based on the Matthews Correlation Coefficient (MCC) and the model inference rate. The score is computed as shown in Equation 16.

$$TOS = \frac{IR}{MCC} \quad (16)$$

TOS refers to Trade-Off Score, IR refers to Inference Rate (seconds per sample), and MCC refers to Matthews Correlation Coefficient.

2.6 Grad-CAM for Visualization

The Grad-CAM [42] technique can be considered as an extension or generalization of the class activation mapping (CAM) [43]. The objective of this study is to explore how the Grad-CAM technique elucidates the multi-label classification outcomes of a model applied to remote sensing images. By utilizing Grad-CAM, this research can visually verify the regions of interest that the network focuses on. This approach ensures that the network accurately identifies and activates the relevant patterns within the image. The expression for the importance of weight of a neuron can be represented as Equation 17.

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k{}_{i,j}} \quad (17)$$

$Z = u \times v$ is the number of pixels in the feature map. Here u and v are spatial dimensions. The Grad-CAM map is subsequently generated by taking a weighted sum of the feature maps, followed by the application of the rectified linear unit (ReLU) activation function that can be seen in Equation 18.

$$M = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (18)$$

The ReLU activation function guarantees the extraction of solely those features that make a positive impact on the class of interest. Consequently, the resulting output is a heatmap that corresponds to the designated class, possessing identical dimensions as the feature map. The Grad-CAM map is subsequently resized to match the dimensions of the input data.

3. Results and Discussions

3.1 Experimental Results

Our experiment results demonstrate the performance of each model in classifying the dataset used for training. The results of the average ensemble, weighted average ensemble, and geometric mean ensemble are presented in Tables 4 and 5. This research tested the first dataset, RSI-CB256, using the EfficientNet B7 Model and the Vision Transformer B16 model. As per Table 5, the dataset has four classes, and both models demonstrated excellent transfer learning performance. The EfficientNet B7 model achieved an accuracy score of 0.9956, while the Vision Transformer B16 model achieved an accuracy score of 0.9919.

Table 4. Performance of EfficientNetB7, ViTB16 and Ensemble learning on RSI-CB256 Dataset

Approach	Prec	Recall	F1-Score	Acc
EfficientNetB7	0.9956	0.9956	0.9956	0.9956
ViTB16	0.9919	0.9919	0.9918	0.9919
Average Ensemble	0.9985	0.9985	0.9985	0.9985
Weigthed Average Ensemble	0.9971	0.9970	0.9970	0.9970
Geometric Mean Ensemble	0.9985	0.9985	0.9985	0.9985

Table 5. Performance of ConvNext Large 384, ViTL16 384 and Ensemble learning on NWPU RESISC45 Dataset

Approach	Prec	Recall	F1-Score	Acc
ConvNext Large 384	0.9644	0.9630	0.9630	0.9630
ViTL16 384	0.9392	0.9381	0.9380	0.9381
Average Ensemble	0.9715	0.9710	0.9710	0.9710
Weigthed Average Ensemble	0.9703	0.9696	0.9696	0.9696
Geometric Mean Ensemble	0.9724	0.9720	0.9720	0.9720

To further enhance the performance of these models, this research employed an ensemble learning approach. The average ensemble and geometric mean ensemble achieved the same accuracy and F1-score of 0.9985,

which is higher than the average ensemble score of 0.9970. These results indicate that EfficientNet B7 and ViTB16 are suitable models for classification.

The following results use the NWPU RESISC45 dataset and the ConvNext Large 384 and ViTL16 384 models. These two models require an input image of 384 x 384 pixels. Table 6 shows the results this research was able to achieve. Overall, the accuracy produced by the ConvNext Large 384 model is 0.9630, while the accuracy produced by the ViTL16 384 model is 0.9381. However, when using the ensemble learning method, the geometric mean ensemble learning achieved an accuracy of 0.9720 and successfully improved the performance. This shows that the remote sensing image classification performance has significantly improved using these two models. To assess the training performance of each model,

Figures 5-8 presents the accuracy and loss curves. These figures indicate that the transfer learning approach is effective, resulting in satisfactory model performance.

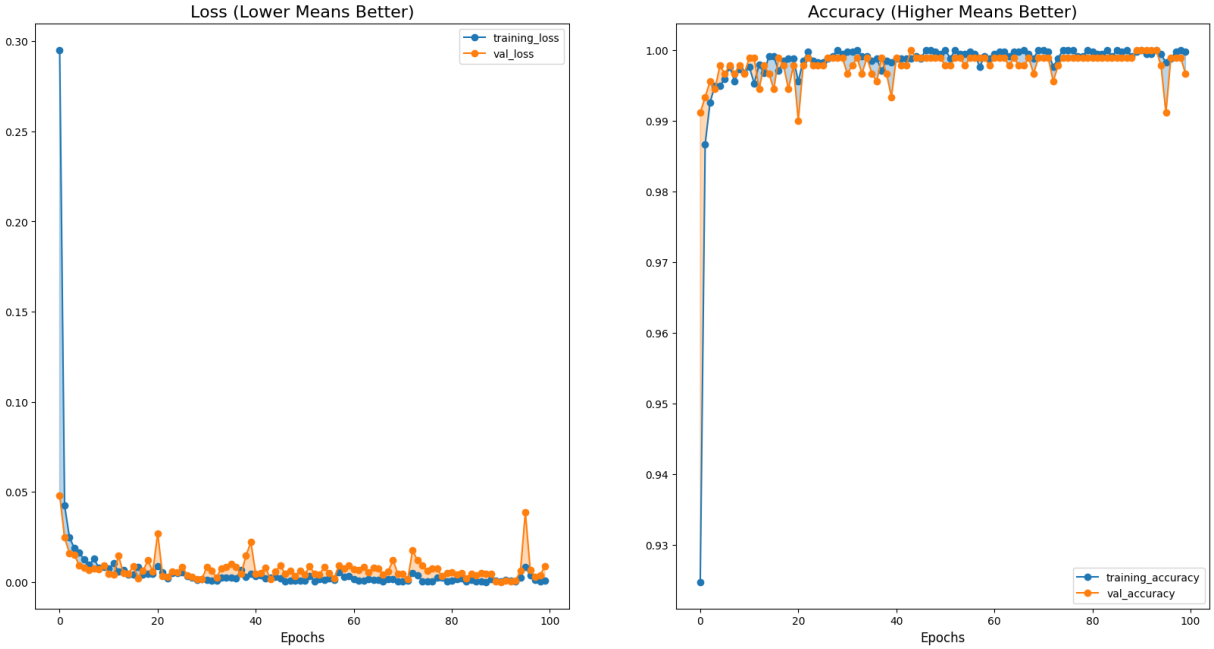


Figure 5. Training curves of EfficientNetB7

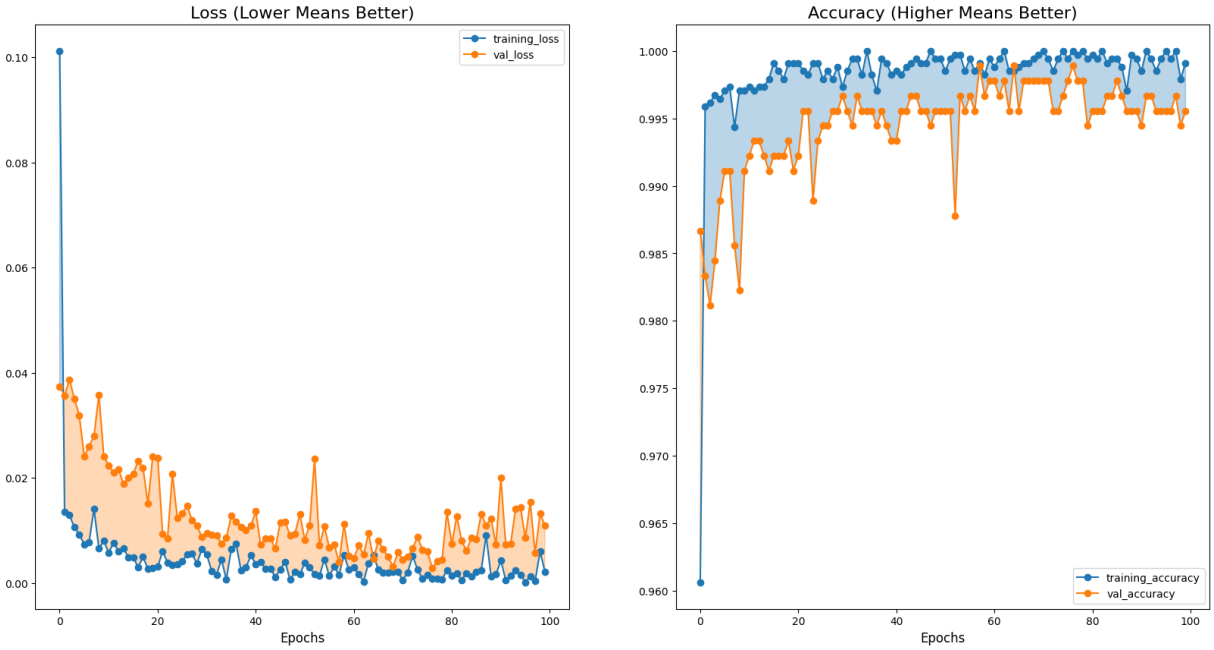


Figure 6. Training curves of ViTB16

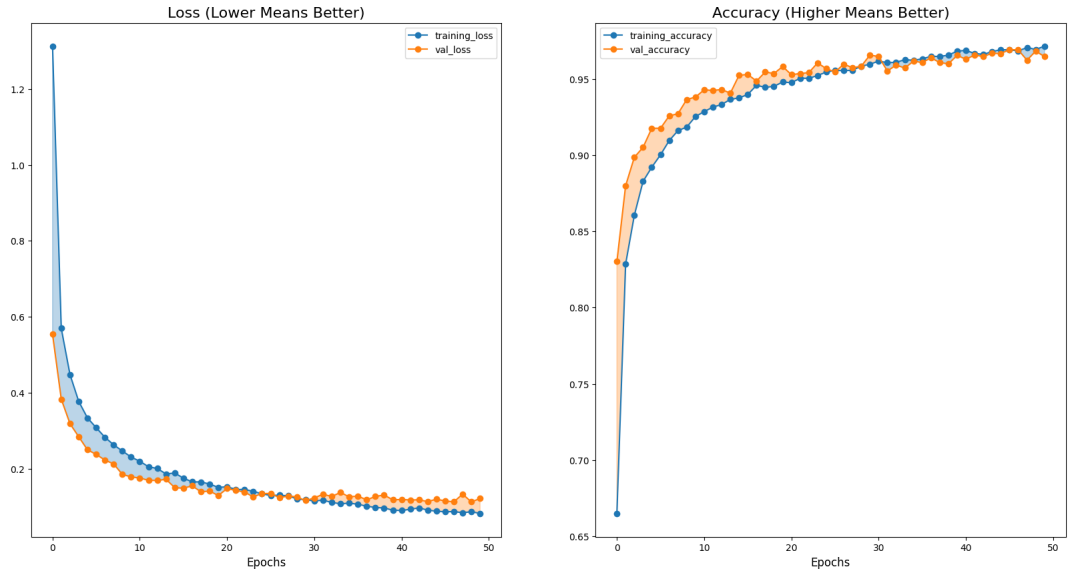


Figure 7. Training curves of ConvNext Large 384

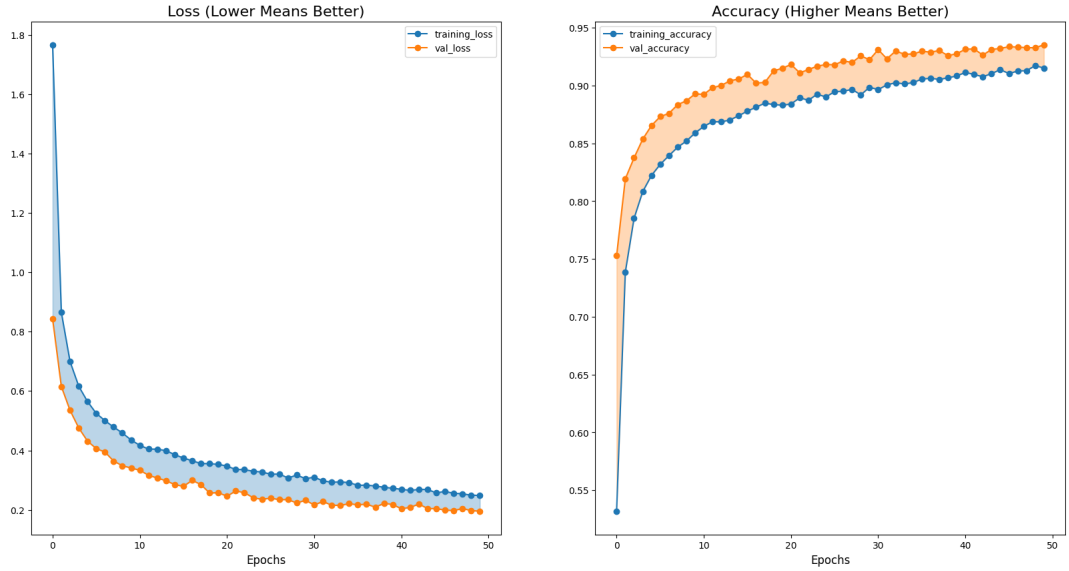


Figure 8. Training curves of ViTL16 384

3.2 Performance Analysis

This study explored transfer learning and ensemble learning strategies for remote sensing image classification across two benchmark datasets namely RSI-CB256 and NWPU RESISC45. These datasets were evaluated using complementary deep learning architectures to assess both classification performance and computational efficiency.

On the RSI-CB256 dataset, which had four classes with clear visual distinctions, EfficientNet-B7 and Vision Transformer B16 (ViT-B16) were evaluated. The limited class complexity allowed both models to achieve satisfactory performance individually. EfficientNet-B7 excelled at capturing local spatial features through its convolutional design, while ViT-B16 leveraged self-attention to model global contextual relationships on the image. When these architectures were integrated into an ensemble, the accuracy improved to 0.9985, demonstrating that the complementarity between CNNs and transformers can produce measurable improvements in predictive robustness even when single-model performance is already high.

On the other hand, the NWPU-RESISC45 dataset presented a more challenging classification scenario with 45 scene classes. To address this complexity, ConvNeXt Large and ViT-L16 were employed at an input resolution of 384×384. As expected, the overall accuracy was lower than on RSI-CB256, reflecting the increased difficulty of the task. However, ensemble learning still improved the classification accuracy to 0.9720, highlighting the value of integrating convolutional and transformer-based representations when class boundaries are ambiguous or overlapping.

Figures 9 and 10 show that computational efficiency was evaluated through a trade-off analysis, comparing inference rate to predictive performance, quantified using the MCC. Single-model architecture consistently

provides the most favorable efficiency balance. On the NWPU RESISC45 dataset, the EfficientNet-based model achieved the lowest trade-off score (0.0601). On the RSI-CB256 dataset, it outperformed all other configurations (0.0082), demonstrating that significant losses in discriminative accuracy did not offset its inference-speed advantage.

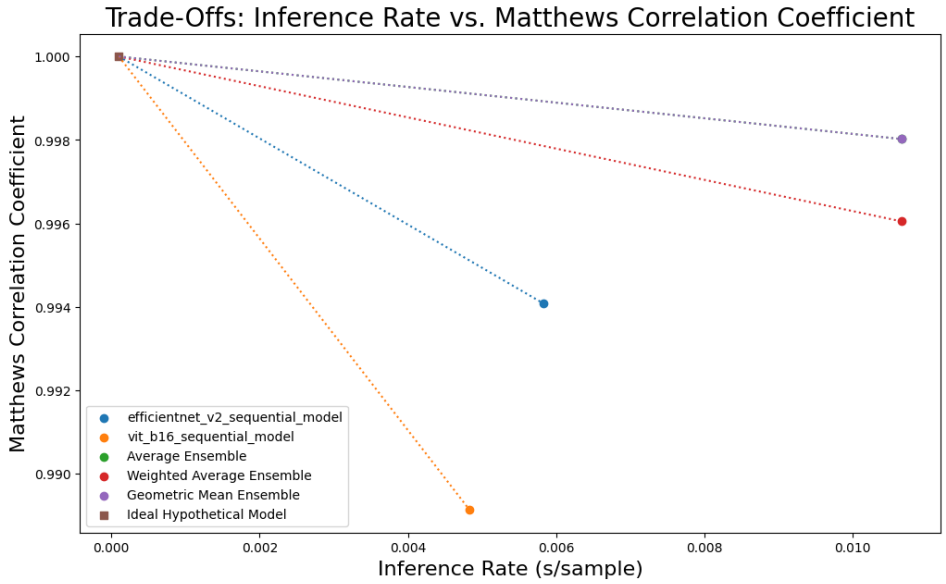


Figure 9. Trade-Offs on RSI-CB256 Dataset

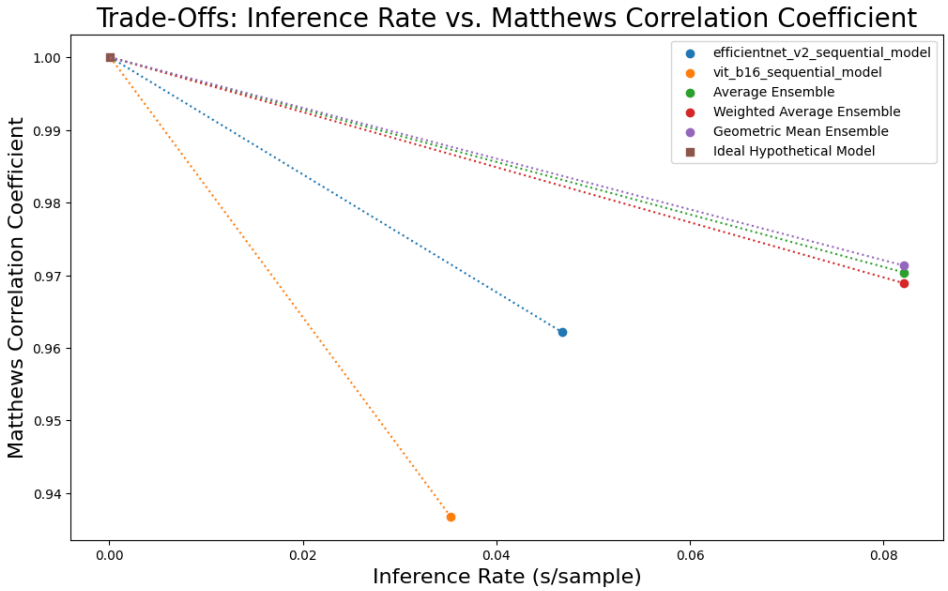


Figure 10. Trade-Offs on NWPU RESISC45 Dataset

Ensemble models achieved marginally higher MCC values, 0.9713 on NWPU RESISC45 and 0.9980 on RSI-CB256, but at the expense of increased inference time due to the aggregation of predictions from multiple models. This performance-cost relationship underscores a clear architectural trade-off: ensembles prioritize predictive accuracy, while single models prioritize deployment efficiency.

Table 6 compares our findings with previous studies. This research conducted a thorough analysis of several related research papers. presents a side-by-side comparison of our current results with those of previous studies. It shows that using transfer learning, GC, and an ensemble improves model performance.

Table 6. Comparison with Previous Research

Previous Studies	Dataset	Approach	Accuracy
Tehsin, et al. [44]	RSI-CB256	CNN with scalable networks for feature extraction.	0.9964
Hasan, et al. [45]	RSI-CB256	CNN-LSTM	0.9880
Zhao, et al. [46]	NWPU RESISC45	Multigranularity Multilevel Feature Ensemble Network (MGML-FENet).	0.9539
Zhai et al. [47]	NWPU RESISC45	Pruning ResNet 34	0.9197
Ours	RSI-CB256 NWPU RESISC45	Transfer Learning and Ensemble Learning.	0.9985 0.9720

3.3 Explainable AI

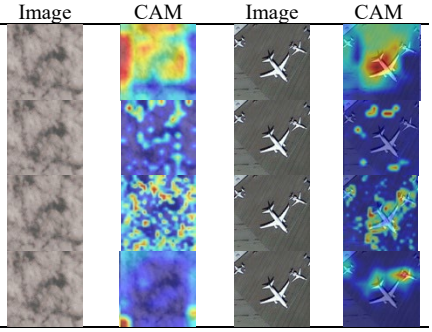
Table 7 presents the results of the Grad-CAM model interpretation. The transfer learning process utilizes pre-trained models ViTB16 224, ConvNextLarge 384, EfficientNetB7, and ViTL16 384. Notably, the ViT16 384 model can visualize the processed input image, as exemplified by a cloud image from the CB256 dataset. The image processing capabilities of the ViT16 384 model enhance the visual explanations of the cloud image, providing more detailed insights. Similarly, the ViTL16 384 model demonstrates the ability to represent the image plane in greater detail when analyzed using samples from the NWPU dataset. The model's image layers are displayed transparently, allowing for a comprehensive understanding of the image.

Other models, such as EfficientNetB7, can still distinguish different parts of the image with more precise visualization explanations. The Grad-CAM results indicate a more robust marking of the cloud region in the image, suggesting that the model effectively detects cloud objects by providing comprehensive visualizations.

In the image plane, the EfficientNet B7 model visualizes the flat parts of the image more clearly, as shown in Table 7. Additionally, the ConvNextL384 model provides more precise and detailed visualizations of both plain and cloud images. While not every part of the image is rendered with Grad-CAM, the consistently highlighted areas appear to correspond to the objects the model identifies during classification.

Table 7 shows that integrating XAI into remote-sensing classification offers substantial advantages. It enhances transparency, interpretability, and trust. By revealing the most influential features and factors contributing to model predictions, XAI facilitates data interpretation and supports more informed and justified decision-making.

Table 7. Visualization Explainable Artificial Intelligence

Model	Cloud from RSI-CB256		Plane from NWPU	
	Original Image	Grad-CAM	Original Image	Grad-CAM
EfficientNetB7				
ViTB16 224				
ViTL16 384				
ConvNext Large 384				

Furthermore, XAI helps identify biases, errors, or model limitations, thereby promoting refinement and optimization to improve performance. This ensures result validation, error correction, and enhanced reliability. Additionally, XAI supports ethical and legal compliance by enabling stakeholders to assess the fairness and accountability of model outputs in remote sensing applications.

4. Conclusions

In this experiment, transfer learning with a pre-trained model achieved promising accuracy for classifying remote-sensing images. However, when compared, these techniques yielded varying results in producing the final classification. These results suggest that the complexity of the dataset influences the behavior of ensembles. The RSI-CB256 dataset, with only four classes and distinct visual differences, led to high agreement among models. Consequently, both the average and geometric ensembles achieved similar performance.

In contrast, the NWPU RESISC45 dataset, with many visually similar classes, led to variance between classifiers. Under these conditions, the geometric mean ensemble performed better because it prioritized consensus among models and minimized the impact of uncertain predictions. XAI visualization via Grad-CAM helps experts select the most suitable pre-trained model for remote sensing image classification. For future research, it is advisable to explore a wider range of ensemble strategies and investigate additional XAI techniques to enhance further model confidence and interpretability in remote sensing classification tasks. The current reliance on only three model combinations per scheme may limit the generalizability of the findings.

Acknowledgements

The authors would like to express their sincere gratitude to Universitas Kristen Satya Wacana for supporting this research through the Internal Research Grant Program under Grant No. 039/SPK-PF/RIK/04/2025.

References

[1] L. Deng and D. Yu, "Deep Learning Methods and Applications," 2014, https://doi.org/10.1561/2000000039.

[2] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: an overview," *Adv. Intell. Syst. Comput.*, vol. 1141, pp. 599–608, 2021, https://doi.org/10.1007/978-981-15-3383-9_54.

[3] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 1–15., https://doi.org/10.48550/arXiv.1706.03762.

[4] S. Paneru and I. Jeelani, "Computer vision applications in construction: Current state, opportunities & challenges," *Autom. Constr.*, vol. 132, p. 103940, Dec. 2021, https://doi.org/10.1016/j.autcon.2021.103940.

[5] J. Lu, L. Tan, and H. Jiang, "Review on Convolutional Neural Network (CNN) Applied to Plant Leaf Disease Classification," Agriculture, vol. 11, no. 8, p. 707, Jul. 2021, https://doi.org/10.3390/agriculture11080707.

[6] F. K. Shaikh, M. Memon, N. A. Mahoto, S. Zeadally, and J. Nebhen, "Artificial Intelligence Best Practices in Smart Agriculture," *IEEE Micro*, pp. 1–1, Oct. 2021, https://doi.org/10.1109/MM.2021.3121279.

[7] K. Bhatia, S. Dhalla, A. Mittal, S. Gupta, A. Gupta, and A. Jindal, "Integrating explainability into deep learning-based models for white blood cells classification," *Comput. Electr. Eng.*, vol. 110, p. 108913, Sep. 2023,

https://doi.org/10.1016/j.compeleceng.2023.108913.
[8] H. Song et al., "Pure data correction enhancing remote sensing image classification with a lightweight ensemble model," Sci. Rep., vol. 15, no. 1, pp. 1–24, Dec. 2025, https://doi.org/10.1038/s41598-025-89735-1.
[9] Z. Zhang *et al.*, "UAV Hyperspectral Remote Sensing Image Classification: A Systematic Review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 3099–3124, 2025, https://doi.org/10.1109/JSTARS.2024.3522318.
[10] P. K. Roy and A. Kumar, "Early prediction of COVID-19 using ensemble of transfer learning," *Comput. Electr. Eng.*, vol. 101, p. 108018, Jul. 2022, https://doi.org/10.1016/j.compeleceng.2022.108018.
[11] A. Abdollahi and B. Pradhan, "Urban vegetation mapping from aerial imagery using explainable AI (XAI)," *Sensors*, vol. 21, no. 14, p. 4738, Jul. 2021, doi: 10.3390/s21144738.
[12] D. K. Sharma, J. Mishra, A. Singh, R. Govil, G. Srivastava, and J. C. W. Lin, "Explainable Artificial Intelligence for Cybersecurity," *Comput. Electr. Eng.*, vol. 103, p. 108356, Oct. 2022, https://doi.org/10.1016/j.compeleceng.2022.108356.
[13] J. Mitchell, N. McLaughlin, and J. Martinez-del-Rincon, "Generating sparse explanations for malicious Android opcode sequences using hierarchical LIME," *Comput. Secur.*, vol. 137, p. 103637, Feb. 2024, https://doi.org/10.1016/j.cose.2023.103637.
[14] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review," *Comput. Biol. Med.*, vol. 166, p. 107555, Nov. 2023, https://doi.org/10.1016/j.compbiomed.2023.107555.
[15] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach," *Expert Syst. Appl.*, vol. 238, p. 121751, Mar. 2024, https://doi.org/10.1016/j.eswa.2023.121751.
[16] C. van Zyl, X. Ye, and R. Naidoo, "Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP," *Appl. Energy*, vol. 353, p. 122079, Jan. 2024, https://doi.org/10.1016/j.apenergy.2023.122079.
[17] R. Bharti, D. Saini, and R. Malik, "A Novel Approach of Transfer Learning for Satellite Image Classification," *Lect. Notes Data Eng. Commun. Technol.*, vol. 90, pp. 379–392, 2022, https://doi.org/10.1007/978-981-16-6289-8_32.
[18] L. Pham, K. Tran, D. Ngo, J. Lampert, and A. Schindler, "Remote sensing image classification using transfer learning-and attention-based deep neural network," p. 19, Oct. 2022, https://doi.org/10.1117/12.2635320.
[19] F. E. Jannat and A. R. Willis, "Improving Classification of Remotely Sensed Images with the Swin Transformer," *Conf. Proc. - IEEE SOUTHEASTCON*, vol. 2022-March, pp. 611–618, 2022, https://doi.org/10.1109/SoutheastCon48659.2022.9764016.
[20] Z. Xia, G. Gan, S. Liu, W. Cao, and J. Cheng, "Global and Multi-Scale Feature Learning for Remote Sensing Scene Classification," *Int. Geosci. Remote Sens. Symp.*, vol. 2022-July, pp. 655–658, 2022, https://doi.org/10.1109/IGARSS46834.2022.9884011.
[21] A. A. Adegun, S. Viriri, and J. R. Tapamo, "Review of deep learning methods for remote sensing satellite images classification: Experimental survey and comparative analysis," J. Big Data, vol. 10, no. 1, 2023, https://doi.org/10.1186/s40537-023-00772-x.
[22] A. Maracani, V. P. Pastore, L. Natale, L. Rosasco, and F. Odone, "In-domain versus out-of-domain transfer learning in plankton image classification," *Sci. Rep.*, vol. 13, no. 1, Jun. 2023, https://doi.org/10.1038/s41598-023-37627-7.
[23] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 103, p. 102520, Dec. 2021, https://doi.org/10.1016/j.jag.2021.102520.
[24] A. Amelio, G. Bonifazi, E. Corradini, D. Ursino, and L. Virgili, "A Multilayer Network-Based Approach to Represent, Explore and Handle Convolutional Neural Networks," Cognitive Computation, vol. 15, no. 1, pp. 61–89, 2023, https://doi.org/10.1007/s12559-022-10084-6.
[25] P. J. Phillips *et al.*, "Four Principles of Explainable Artificial Intelligence", https://doi.org/10.6028/NIST.IR.8312.
[26] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021, https://doi.org/10.1016/j.inffus.2021.05.009.
[27] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, https://doi.org/10.1109/JPROC.2020.3004555.
[28] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: A friendly introduction," J. Big Data, vol. 9, no. 1, pp. 1–19, 2022, https://doi.org/10.1186/s40537-022-00652-w.
[29] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, Oct. 2022, https://doi.org/10.1016/j.engappai.2022.105151.
[30] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 2, pp. 757–774, Feb. 2023, https://doi.org/10.1016/j.jksuci.2023.01.014.
[31] H. Yong, J. Huang, X. Hua, and L. Zhang, "Gradient Centralization: A New Optimization Technique for Deep Neural Networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12346 LNCS, pp. 635 – 652, 2020, https://doi.org/10.1007/978-3-030-58452-8_37.
[32] M. I. Hossen, M. Awrangjeb, S. Pan, and A. Al Mamun, "Transfer learning in agriculture: A review," Artif. Intell. Rev., vol. 58, no. 4, pp. 1–37, 2025, https://doi.org/10.1007/s10462-024-11081-x.
[33] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), vol. 97, 2019, pp. 6105–6114, https://doi.org/10.48550/arXiv.1905.11946.
[34] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 11966–11976, Jan. 2022, https://doi.org/10.1109/CVPR52688.2022.01167.
[35] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, https://doi.org/10.48550/arxiv.2010.11929.
[36] Z. Hong, Z. Yang, H. Wang, D. Li, W. Nai, and Y. Xing, "The weighted average ensemble learning based on polar bear algorithm with T-distribution parameters," pp. 1902–1905, 2020, https://doi.org/10.1109/ITAIC49862.2020.9338792.
[37] S. Wu, J. Li, and W. Ding, "A geometric framework for multiclass ensemble classifiers," Mach. Learn., vol. 112, no. 12, pp. 4929–4958, Dec. 2023, https://doi.org/10.1007/s10994-023-06406-w.
[38] M. Reda, "Satellite Image Classification," Kaggle. Accessed: Dec. 25, 2023. [Online]. Available: https://www.kaggle.com/datasets/mahmoudreda55/satellite-image-classification.
[39] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, https://doi.org/10.1109/JPROC.2017.2675998.
[40] P. Christen, D. J. Hand, and N. Kirielle, "A review of the F-measure: Its history, properties, criticism, and alternatives," *ACM Computing Surveys*, vol. 56, no. 3, Art. no. 73, pp. 1–24, Oct. 2023, https://doi.org/10.1145/3606367.
[41] K. A. Hambarde and H. Proença, "Information Retrieval: Recent Advances and Beyond," IEEE Access, vol. 11, pp. 76581–76604, Jul. 2023, https:doi.org/10.1109/ACCESS.2023.3295776.
[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2016, https://doi.org/10.1007/s11263-019-01228-7.
[43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba,

"Learning Deep Features for Discriminative Localization," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 2921–2929, Dec. 2016, https://doi.org/10.1109/CVPR.2016.319.

[44] S. Tehsin, S. Kausar, A. Jameel, M. Humayun, and D. K. Almofarreh, "Satellite Image Categorization Using Scalable Deep Learning," *Appl. Sci.*, vol. 13, no. 8, p. 5108, Apr. 2023, https://doi.org/10.3390/app13085108.

[45] N. Hasan and M. S. Islam, "A lightweight CNN-LSTM hybrid model for land cover classification in satellite imagery," Geocarto Int., vol. 40, no. 1, 2025, https://doi.org/10.1080/10106049.2025.2521828.

[46] Q. Zhao, S. Lyu, Y. Li, Y. Ma, and L. Chen, "MGML: Multigranularity Multilevel Feature Ensemble Network for Remote Sensing Scene Classification," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 5, pp. 2308–2322, May 2023, https://doi.org/10.1109/TNNLS.2021.3106391.

[47] D. Zhai *et al.*, "Multi-dimensional spatial pruning for remote sensing image scene classification," *Digit. Signal Process.*, vol. 158, p. 104979, Mar. 2025, https://doi.org/10.1016/J.DSP.2025.104979.