Journal of Agricultural Sciences
(Tarim Bilimleri Dergisi)

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

Lightweight Transformer Model for Agricultural Land Use and Land Cover Classification

Kemal Celik[a]

[a]*Department of Geomatics Engineering, Gumushane University, 29100, Gumushane, TURKEY*

ARTICLE INFO
Research Article
Corresponding Author: Kemal Celik, E-mail: gumuscelik@hotmail.com
Received: 22 January 2025 / Revised: 17 April 2025 / Accepted: 12 May 2025 / Online: 30 September 2025

Cite this article
Kemal Celik (2025). Lightweight Transformer Model for Agricultural Land Use and Land Cover Classification. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 31(4):941-959. DOI: 10.15832/ankutbd.1624812

ABSTRACT
Observing agricultural land use via remote sensing images is essential for ensuring food security, estimating yields and planning efficient exports nonetheless precise classification continues to be difficult because of the varied and evolving characteristics of agricultural environments. This research aims to evaluate and optimize advanced deep learning architectures particularly Vision Transformer (ViT) models for agricultural land-use classification tasks. Specifically, we employed ViT-Base-16 and other lightweight models DeiT-Tiny and EfficientNet-B0 applying techniques such as model layer compression and advanced data augmentation CutMix and Cutout to achieve high accuracy while significantly reducing computational complexity. Evaluation was performed using three benchmark remote sensing datasets EuroSAT, NWPU-RESISC45 and SIRI-WHU which include diverse spatial resolutions and agricultural classes relevant for practical monitoring. Findings indicate that the optimized ViT algorithm is highly effective in recognizing global spatial connections, consistently achieving remarkable classification accuracy exceeding 99% on a newly assembled dataset containing around 200 samples of Google Earth imagery. Furthermore, for the first time in agricultural image classification compressing the ViT-Base model by pruning 50% of its layers significantly reduced complexity maintaining competitive accuracy 97.9% on SIRI-WHU. The resulting models are particularly suitable for deployment on devices with limited computational resources supporting real-world operational agricultural monitoring systems. This study emphasizes the revolutionary possibilities and practical use of optimized transformer-based models that offer scalable and efficient solutions specifically designed for precision agriculture applications.

Keywords: Agriculture, Data augmentation, Deep learning, Vision transformers, Land use

1. Introduction

Agricultural land use monitoring is crucial, facilitating tasks including food security management, yield prediction, crop estimation and export planning. Among these accurately identifying fields during cultivation cycles is a key challenge (Papageorgiou et al. 2011; Shastry et al. 2017). Satellite based solutions provide several advantages such as global accessibility at no expense comprehensive geographical coverage and high levels of temporal resolution which facilitate ongoing monitoring capabilities throughout the entire year (Huete et al. 2002). Leveraging the multispectral properties of satellite images various indices have been developed to distinguish crop areas from non-crop regions (Zhang et al. 2018). However, these indices often require extensive domain expertise to design and remain vulnerable to challenging environmental conditions. Traditional image classifiers such as VGG and InceptionNet perform well on RGB images of natural objects but tend to struggle with spectral data due to their limited adaptation to such inputs (Slavkovikj et al. 2015). Domain-specific classifiers such as Support Vector Machines and Convolutional Neural Networks (CNNs) have been designed to address these challenges, but they frequently fall short in capturing the complex relationships between spectral channels and neglecting the crucial temporal aspects necessary for accurate agricultural monitoring (Albert et al. 2017; Park et al. 2018).

Numerous techniques have been proposed in the literature for classifying land cover using remote sensing (RS) images, primarily categorized into supervised and unsupervised learning approaches (Khan & Basalamah 2023). Methods of unsupervised learning typically employ clustering algorithms such as K-means and fuzzy C-means though these approaches tend to be inefficient and not well-suited for handling the significant volume of labelled RS images currently available (Zhao et al. 2023). Conversely conventional supervised learning techniques typically employ predefined features in conjunction with machine learning algorithms (Guo et al. 2023). These handcrafted features are designed using prior knowledge focusing on attributes such as texture, shape and key points. However, in complex scenes, identifying and extracting the most discriminative features remains a significant challenge (Temenos et al. 2023). To address this feature selection has become a critical focus in pattern recognition with various techniques developed to enhance classification efficiency (Nie et al. 2024). These feature selection methods aim to eliminate irrelevant information from the original feature space thereby reducing computational complexity and improving overall performance (Hamza et al. 2023).

Deep learning has achieved remarkable success across various domains, particularly in RS and object classification tasks. Its popularity stems from its ability to handle large datasets efficiently while delivering superior performance (Miotto et al. 2017). CNN’s have become the leading deep learning models capable of extracting hierarchical and abstract image features through a sequential layer-by-layer process (Basu et al. 2015). CNNs are widely applied in fields such as healthcare, action recognition, satellite imaging, fraud detection and more, showcasing their versatility. A typical architecture includes multiple layers, such as convolutional, pooling, ReLU activation, batch normalization, fully connected (FC) and softmax layers each contributing to feature extraction and classification (Gandhar et al. 2024). Recent advancements have leveraged pretrained deep learning models to enhance land cover classification using RS imagery (Singh et al. 2022; Albarakati et al. 2024). Proposed a novel CNN-based architecture integrating multilayer perceptrons and vision transformers specifically designed for RS image land cover classification highlighting the evolving role of deep learning in the field (Hamza et al. 2024).

Recent studies have increasingly explored the application of deep learning for agricultural yield classification using RS imagery (Ahmad et al. 2022; Nguyen et al. 2020). These approaches often involve a complex process that combines features from multiple models and then selects the most relevant ones for classification (Attri et al. 2023; Joshi et al. 2023). However, this feature-level fusion can be computationally intensive and inefficient (Thirumaladevi et al. 2023a). More optimized methodologies such as network-level fusion have been suggested. Network level fusion streamlines the process by integrating models at a structural level enabling more efficient feature extraction and selection while maintaining high classification accuracy (Hamza et al. 2023; Vohra & Tiwari 2023). This approach highlights the growing emphasis on optimization and scalability in agricultural monitoring using deep learning (Zahra et al. 2024).

This study presents a comprehensive evaluation of a ViT based model for the classification of agricultural imagery using remote sensing data. Monitoring agricultural land use is crucial for ensuring food security, predicting yields and planning exports effectively. However, it remains challenging due to the diverse and dynamic characteristics of agricultural landscapes. To overcome limitations related to accuracy and computational efficiency in conventional deep learning methods, agricultural images are divided into smaller patches, flattened and transformed into sequential representations, augmented with positional embeddings to preserve critical spatial information (Dosovitskiy et al. 2020). Subsequently multiple multi-head attention layers extract meaningful features capable of capturing global spatial dependencies and fine-grained details in agricultural fields. The evaluation performed on three publicly available datasets EuroSAT, NWPU-RESISC45, and SIRI-WHU demonstrates that ViT-Base-16 (384x384) achieves outstanding classification accuracies 98.8% on NWPU, 99.7% on SIRI-WHU and 99.5% on EuroSAT highlighting its suitability for detailed agricultural monitoring. Additionally, lightweight alternatives such as DeiT-Tiny and EfficientNet-B0 achieve competitive performance 94.3% and 94.1% accuracy on NWPU, respectively offering computationally efficient solutions suitable for resource constrained environments. A pruned ViT-Base model (224x224) achieves an accuracy of 97.9% on the SIRI-WHU dataset significantly reducing the model's complexity by 50% with minimal loss in performance. Additionally, advanced data augmentation techniques such as CutMix and Cutout further improve the model's robustness and generalizability. By tailoring deep learning architectures to specific dataset challenges, this research provides scalable, efficient, and highly accurate classification methods contributing novel insights to precision agriculture and land management applications.

2. Material and Methods

2.1. Dataset

This study employs three publicly accessible datasets EuroSAT, NWPU-RESISC45 and SIRI-WHU for classification tasks in the remote sensing domain. All the datasets consist of RGB image data and are intended for research purposes. A brief description of each dataset follows.

The EuroSAT dataset serves as a comprehensive benchmark for the classification of land use land cover and is based on freely accessible Sentinel-2 satellite imagery obtained from the ESA's Copernicus program (Helber et al. 2019). Featuring 27,000 labeled and georeferenced 64x64 image patches across 10 diverse classes such as residential areas, forests, water bodies and crops. It spans 13 spectral bands including visible and infrared wavelengths. Designed to facilitate applications like environmental monitoring, urban planning and change detection. The dataset offers high intra class variance by incorporating data from 34 European countries over various seasons. Achieving a benchmark accuracy of 98.57% using deep learning models like ResNet-50. EuroSAT enables advanced multispectral analysis while remaining open for both academic and practical use making it a pivotal resource for Earth observation research.

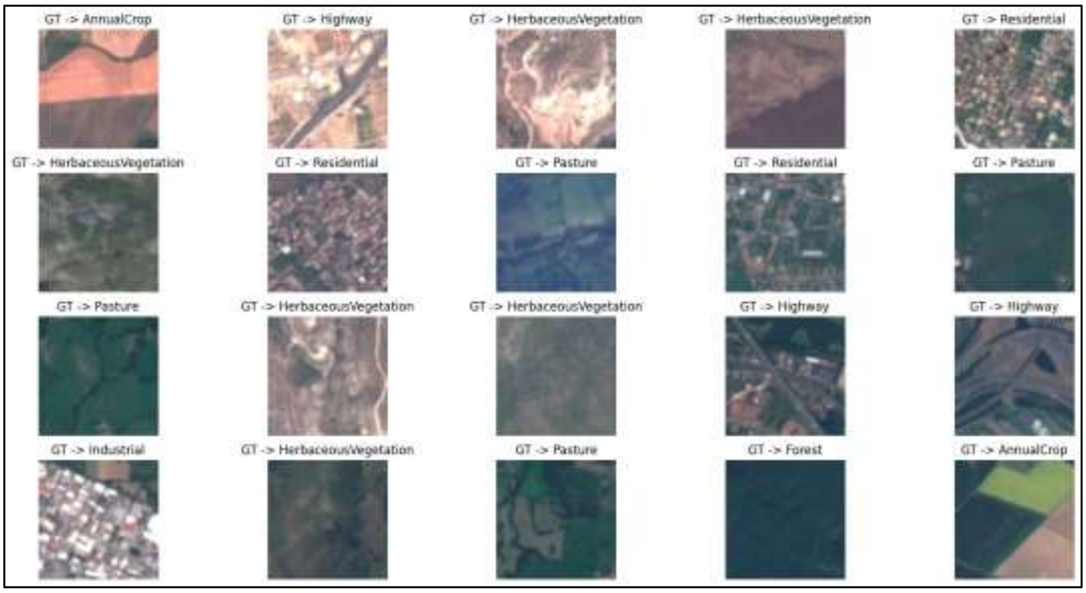

Figure 1- EuroSat Dataset

The SIRI-WHU dataset created by Wuhan University's Intelligent Data Extraction and Analysis Group, consists of 2,400 high-resolution remote sensing images each measuring 2 meters in spatial resolution and having dimensions of 200x200 pixels (Zhao et al. 2016). The dataset consists of 12 categories including agriculture, harbor, industrial and residential. Each category comprising 200 images that were sourced from Google Earth. While the dataset effectively supports land use classification research its geographic focus on urban areas in China and limited diversity of classes make it more suitable for specific applications rather than comprehensive global studies.

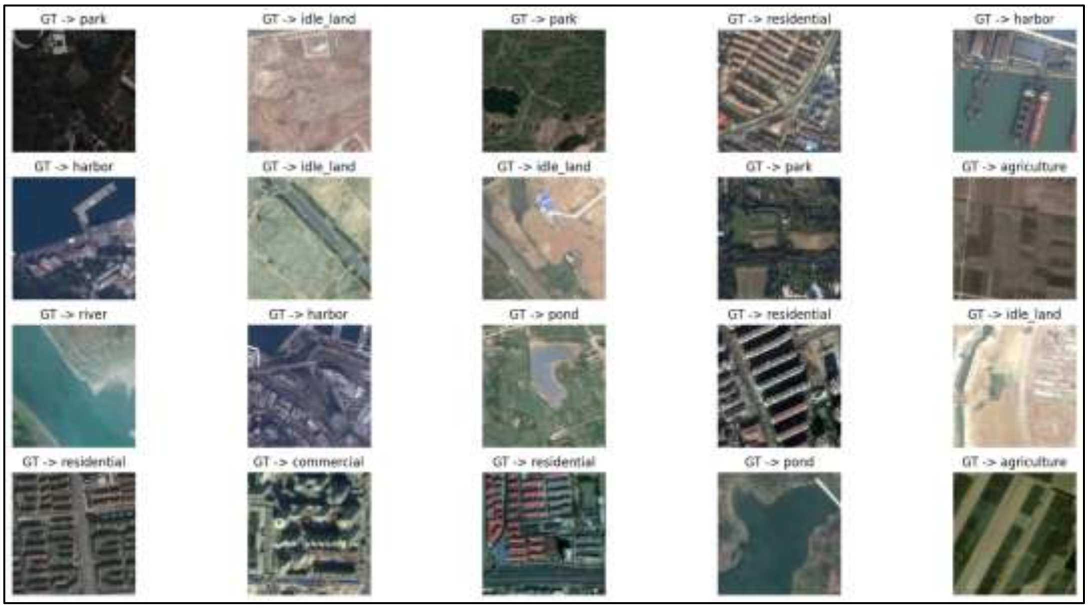

Figure 2- SIRI-WHU Dataset

Northwestern Polytechnical University launched the NWPU-RESISC45 dataset to serve as a standard for evaluating remote sensing image scene classification (Cheng et al. 2017). The dataset comprises 31,500 images which are categorised into 45 scene classes with each class containing 700 images. The system was intended to overcome the limitations of smaller data sets by incorporating various translations different spatial resolutions, multiple object positions, diverse lighting conditions and occlusions. Images sized 256x256 pixels originated from Google Earth encompassing over 100 countries and range from a spatial resolution of approximately 0.2 to 30 meters. The dataset is designed to aid in the creation and assessment of sophisticated algorithms that rely on data particularly advanced deep learning techniques by providing a high level of diversity within each category and similarity between categories. This dataset has established a new benchmark for scale and intricacy in the classification of images from remote sensing.

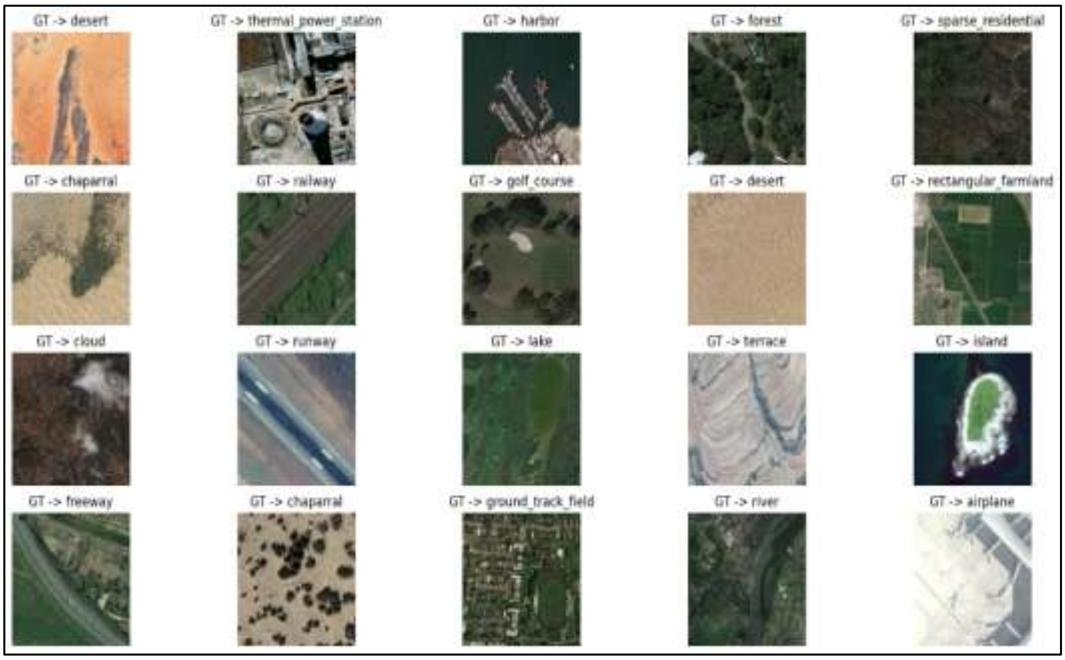

Figure 3- NWPU-RESISC45 Dataset

2.2. Vision transformer

The ViT introduces the Transformer model originally created for NLP into the realm of image classification. This innovative approach diverges from traditional CNNs which rely on localized filters to analyse image regions. ViT models employ self-attention mechanisms to identify long-range dependencies within an image thereby gaining a more detailed comprehension of visual information (Khan et al. 2021). The ViT model builds upon the established vanilla Transformer architecture capitalizing on its demonstrated effectiveness in applications such as machine translation and other natural language processing tasks (Thakur et al. 2023). In contrast to recurrent network-based models, the Transformer's encoder-decoder architecture handles sequential data in a parallel manner (Ranftl et al. 2021). At the core of its effectiveness is the self-attention mechanism which effectively represents and retains relationships within sequences enabling precise analysis and contextual comprehension. In order to adapt the Transformer for vision tasks ViT views images as sequences similar to how words are processed in sentences (Zhai et al. 2021). The image is segmented into numerous non overlapping sections and each one is mapped into a vector space to maintain its distinctive features. These embeddings are then fed into a Transformer encoder thereby eliminating the requirement for conventional CNN operations and still enabling efficient and scalable image analysis (Li & Li 2022).

The ViT was presented as a novel approach to boost the standard Transformer architecture's capabilities in image classification tasks. Its main goal is to widen the model's use to different types of data formats beyond text without having to use architectures specifically designed for particular data types (Zuo et al. 2022). ViT accomplishes this by utilising the Transformer's encoder component to map sequences of image segments to semantic labels. Unlike traditional CNNs that rely on localized receptive fields. ViT model uses its attention mechanism to focus on multiple parts of an image and synthesize information from the entire image to achieve a comprehensive global perspective (Bazi et al. 2021).

The ViT's end-to-end architecture was created with a focus on both simplicity and its ability to be effective. It consists of three main components: an embedding layer, a Transformer encoder and a classification head. The process begins by dividing an image from the dataset designated as X, into non-overlapping patches (Khan et al. 2021). Each patch serves as an individual token for the Transformer. For an image of dimensions c×w×h, where c is the number of channels, h is the height and w is the width. Patches of size c×p×p are extracted (Zhai et al. 2021). This process transforms the image into a sequence of n patches, where $n = \frac{hw}{p^2}$. Smaller patch sizes such as 16×16, result in longer sequences while larger patches like 32×32 yield shorter ones (Han et al. 2020).

ViT models patches as individual tokens mirroring how transformers process text in natural language processing. Each token is embedded into a vector representation that captures its features, and these embeddings are processed by the Transformer's encoder (Xia et al. 2022). The encoder uses self-attention mechanisms to investigate connections between patches allowing the model to detect patterns and characteristics throughout the entire image. This approach contrasts with CNNs where information is processed hierarchically and often localized in nature. Through its sequence-based design ViT provides a powerful framework for comprehensive image analysis and classification (Alzahrani & Alsaade 2023).

2.3. Vision transformer variants

Variants of the ViT have been developed to study how model size impacts classification precision encompassing the ViT-Base, ViT-Large and ViT-Huge. The different configurations vary in several architectural components including the number of encoder layers the hidden dimensionality, the amount of attention heads within the Multi-Head Self-Attention (MSA) blocks and the size of the MLP classifier (Maurício et al. 2023).

Table 1- Model parameters and layers of ViT

Model	*Layers*	*Hidden Size*	*MLP Size*	*Heads*	*Parameters*
ViT-Base	12	768	3072	12	86 m
ViT-Large	24	1024	4096	16	307 m
ViT-Huge	32	1280	5120	16	632 m

The variants shown here illustrate the ability of the ViT framework to be adapted and enlarged. While larger models such as ViT-Huge excel in capturing complex relationships and achieving superior accuracy. They also demand more computational resources. By comparing these versions researchers can assess the trade-offs between computational cost and classification effectiveness tailoring the model to specific tasks and constraints (Gong et al. 2023).

Research demonstrates that Vision Transformers of varying sizes display higher accuracy rates with larger model configurations. Choosing smaller patch sizes leads to an increase in the sequence length (n) thereby boosting the model's overall performance. It's noteworthy that the attention heads in the initial layers of ViTs can focus on image regions that are distant from one another. The model's capacity for this capability becomes even more evident when its depth is increased. This behaviour contrasts sharply with CNN-based models where earlier layers primarily capture localized features, and global features are only recognized in the deeper layers. In ViTs the ability to attend to both local and distant regions from the outset proves invaluable for extracting critical features needed for accurate classification. This unique characteristic underscores the model's effectiveness in handling complex visual data (Reedha et al. 2022).

2.4. Linear embedding layer

Transformation facilitated by a trainable embedding matrix E. This step transforms the flattened image patches into feature representations compatible with the model. Additionally, a learnable classification token v_{class} is prepended to the sequence serving as a marker for the classification process. The Transformer views these embedded patches as separate tokens without any inherent understanding of their spatial positioning within the image. To preserve the original spatial structure of the image positional encoding, denoted as E_{pos}, is added to the patch embeddings. This encoding explicitly represents the relative positions of the patches within the sequence. The final embedded sequence, including the classification token is expressed as:

$$z_0 = [v_{class};\ x_1E; x_2E; \dots; x_nE] + E_{pos}, \quad E \in \mathbb{R}^{(p^2c)\times d}, \ \ E_{pos} \in \mathbb{R}^{(n+1)\times d} \tag{1}$$

Research has demonstrated that both 1-D and 2-D positional encodings perform similarly. Consequently, the simpler 1-D encoding is typically employed to effectively capture positional relationships among the patches in the sequence (Vaswani et al. 2017).

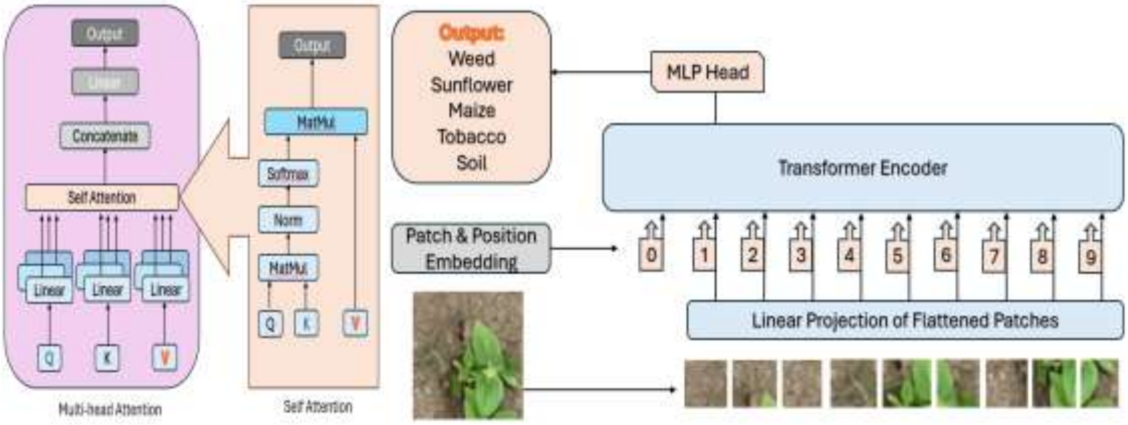

Figure 4- Architecture of the ViT model (Celik et al. 2024)

2.5. Vision transformer encoder

The sequence of embedded patches, denoted as z_0, is fed into the Transformer encoder. The encoder consists of L identical layers each designed to refine the representation of the input sequence progressively (Bazi et al. 2021). Each layer's composition consists of two fundamental components. A multi-head self-attention block which is outlined in Equation 2, and a fully connected feed-forward dense block as specified in Equation 3, this block enhances feature representation through two dense layers that are divided by an activation function as illustrated in Equation 4. The two subcomponents are enhanced with residual skip connections and are preceded by a normalization layer (LN) to stabilize the training process (He et al. 2016a ; Chen et al. 2021).

The operations in each layer can be expressed as:

$$z_l' = \text{MSA}\ (\text{LN}(z_{l-1})) + z_{l-1}, \quad l = 1,\dots,L \tag{2}$$

$$z_l = \text{MSA}\ (\text{LN}(z_{l-1})) + z_l'\ , \quad l = 1,\dots,L \tag{3}$$

Where; z_l' represents the output after applying multi-head self-attention and z_l is the final output of the layer after the feed-forward operation. This layered design allows the encoder to iteratively refine the patch embeddings learning both local and global relationships essential for downstream tasks like classification. In the last layer of the encoder the initial element of the sequence z_0^L, is extracted and passed to an external classification head to predict the class label. This method ensures that the classification token which integrates information from all patches throughout the encoding process is effectively used for the prediction task (Shin et al. 2023).

$$y = \text{LN}(z_l^0\ ) \tag{4}$$

The primary purpose of the encoder's MSA block is to assess the relative importance of each patch embedding in relation to its counterparts within the sequence. The MSA block comprises four key components. A linear layer that translates inputs into query, key and value vectors, a self-attention layer which computes attention scores, a concatenation layer employed to combine outputs from multiple attention heads and concluding linear layer for dimensionality reduction integration. The MSA block's structure enables it to identify complex relationships between image patches thus allowing the model to concentrate on the most pertinent features for subsequent tasks (Zhai et al. 2021; Beyer et al. 2022).

$$[\text{Q, K, V}] = zU_{QKV}, U_{QKV} \in\ \mathbb{R}^{d\times 3D_K} \tag{5}$$

At its core the attention mechanism calculates the importance of each component in a sequence by computing attention weights representing a weighted total of all values in the sequence z. In a self-attention head these weights are acquired through a query-key-value dot-product computation. Each input element is transformed into three vectors. Q (query), K (key), and V (value), by multiplying the element with three distinct learned matrices U_Q, U_K, and U_V (Equation 5) (Bazi et al. 2021).

$$\text{A} = \text{softmax}\left(\frac{QK^T}{\sqrt{D_K}}\right), A \in\ \mathbb{R}^{n\times n} \tag{6}$$

To evaluate the relevance of one element in the sequence to the others, the dot product between its Q vector and the K vectors of all other elements is calculated. These results quantify the relative importance of different patches in the sequence. The outputs of this operation are scaled by the dimensionality of the key vectors D_K, to stabilize gradients, and then passed through a softmax function to convert them into probabilities (Equation 6). Finally, these probabilities are used as weights to compute the weighted sum of the V vectors, producing the attention scores that highlight the most important elements in the sequence (Equation 7) (Bazi et al. 2021; Shin et al. 2023). The full operation can be summarized as follows:

$$\text{SA(z)} = \text{A.V} \tag{7}$$

This process enables the self-attention block to focus on the most relevant regions in the sequence making it a powerful mechanism for identifying and modelling relationships between patches in the input data. The MSA block enhances the standard self-attention mechanism by computing scaled dot-product attention across multiple attention heads. Instead of relying on a single set of Query, Key and Value, the MSA block uses multiple sets to capture diverse relationships within the input data (Suh et al. 2018; Han et al. 2020). Each attention head performs the self-attention operation independently processing different subspaces of the input representation. The outputs of all attention heads are next combined into a single vector which is then transformed through a feed-forward layer utilizing learnable weights W to achieve the desired output dimension (Thakur et al. 2023). This enables the model to aggregate information from multiple perspectives effectively. The entire MSA operation can be expressed mathematically as:

$$\text{MSA(z)} = \text{Concat}\ (\text{SA}_1(z), \text{SA}_2(z), \dots, \text{SA}_h(z))\ W,\ \ W \in\ \mathbb{R}^{hD_K\times D} \tag{8}$$

Where; h represents the number of attention heads, D_K is the dimensionality of each head's output and D is the total output dimension (Zuo et al. 2022). This multi-headed approach allows the MSA block to learn richer and more nuanced representations by attending to different features of the input sequence simultaneously.

2.6. Data augmentation strategy

Data augmentation is a powerful, yet straightforward technique used to enhance the size and variety of training datasets. This is particularly crucial when large annotated datasets are unavailable. Existing data is modified using various manipulation techniques to generate new training samples thereby preserving the original class labels intact. Reducing overfitting and enhancing the model's robustness and ability to generalise effectively are the primary goals of this method.

Standard data augmentation techniques involve applying various geometric transformations like rotation, scaling, cropping, shifting and flipping, either on their own or in conjunction with one another (Jackson et al. 2018). These operations generate multiple versions of the original images while maintaining their primary characteristics (Bowles et al. 2018). Adjustments to brightness, contrast or colour saturation, like colour-space transformations. Also enhance the diversity of datasets. These methods are expanded upon by neural style transfer which alters fundamental image features such as texture and then applies the style of one image to another while preserving the target image's underlying meaning.

Generative models like GANs are used in more sophisticated augmentation methods to generate synthetic samples that closely replicate the original data distribution. This approach effectively increases the dataset's richness by creating entirely new, yet realistic, examples. These diverse augmentation strategies collectively strengthen the model's ability to generalize across unseen data making data augmentation a fundamental component of modern machine learning workflows. Recent advancements in data augmentation have introduced more sophisticated techniques such as Cutout, Mixup and CutMix, to generate challenging and diverse training samples. These methods aim to push models beyond relying on specific features, promoting a deeper understanding of the data.

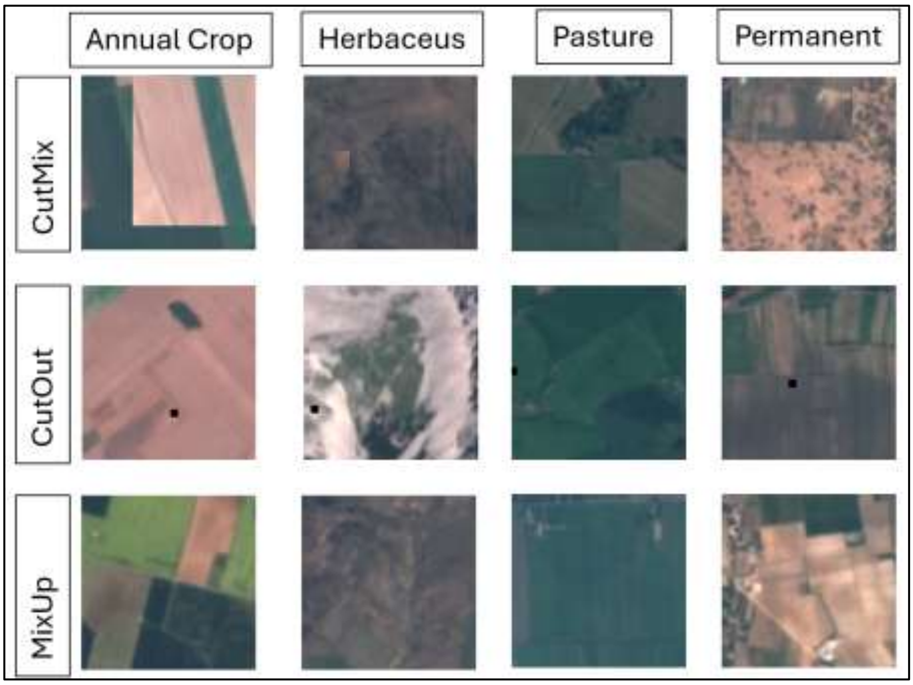

Figure 5- Augmented dataset image examples

Replacing a specified area of an image with either black pixels or random noise is a process known as cutout. This approach allows the model to learn from the image's broader context rather than focusing on specific visual details. A major limitation of Cutout is that it can obscure key aspects of an image, potentially leading to a loss of essential information (DeVries & Taylor 2017). To mitigate the risk of losing critical information, CutMix replaces a section of one image with a piece from another image in the dataset thereby introducing diversity in the training data (Yun et al. 2019).

Mixup takes a different approach by blending two images and their corresponding labels through linear interpolation (Zhang et al. 2017). This technique creates new training instances by combining samples $\chi_i$and χ_j along with their respective labels ψ_i and ψ_j. The interpolated dataset is expressed mathematically as:

$$\chi = \lambda\chi_i + (1-\lambda)\chi_j,\ \psi = \lambda\psi_i + (1-\lambda)\psi_j \quad (9)$$

Where; λ is a mixing factor sampled from the interval [0,1]. Both CutMix and Mixup dynamically alter the ground truth labels to reflect the transformations applied, encouraging the model to generalize better to unseen data and learn more robust feature representations. Examples of the dataset are illustrated in Figure 3, which demonstrates the application of Cutout, Mixup, and CutMix techniques. Selecting the most effective augmentation technique typically demands a customized hands on approach for each individual task. Recent data augmentation developments aim to automate the identification of optimal transformations minimising the requirement for human participation and facilitating more efficient model training (Jackson et al. 2018; Cubuk et al. 2019).

2.7. Network Compression

Transformers are complex neural network architectures that comprise multiple layers, numerous attention heads and millions of parameters. The ViT-Base model is comprised of over 80 million parameters as illustrated in Table 1 (Suravarapu & Patil 2023). Large-scale language models frequently deliver outstanding results, but their substantial memory needs and high computational complexity pose significant challenges for practical use and make them vulnerable to overfitting. To mitigate these challenges techniques for compressing models are used to produce more efficient versions without sacrificing accuracy. Two common strategies are knowledge distillation and model pruning. This process involves transferring the knowledge of a well-trained 'teacher' model to a smaller 'student' model enabling the student model to learn under close supervision how to replicate the teacher's responses. Meanwhile pruning reduces the model size by removing unimportant or redundant parameters. This can be achieved through weight pruning, which eliminates less significant weights or weight quantization which compresses weights by representing them with fewer bits. Table 2 shows compression algorithm.

Table 2- Network Compression Algorithm

Algorithm: Vision Transformer Network Pruning	
1.	Number of batchsize: 64, Optimizer: Adam, Learning Rate: 3e-4, Iteration Number: 10, Input: Training Images (224 px), epoch 30.
2.	Increased batch from the training set. Another set of augmented images is generated using a specific augmentation technique. The model is then trained on the original images and the newly generated augmented images, with the cross-entropy loss being optimised. The loss is subsequently backpropagated, and the model's parameters are updated.
3.	Classify test images
4.	Output: Predicted labels of test images.

In this paper ViT-Base model, with its numerous attention heads and deep encoder layers contains a level of redundancy that makes it suitable for optimization. To exploit this, we propose a method of gradually pruning the encoder layers, creating smaller versions of the model with varying depths. This approach strikes a balance between model size and performance aiming to identify the most compact architecture that still achieves strong classification accuracy. Experiments show that pruning up to 50% of the network maintains competitive results, with algorithm detailing the steps involved in training and compressing the model (Hinton et al. 2015).

3. Experimental Results

Three separate experiments were conducted to assess the ViT in different scenarios. The initial trial focused on examining how various data enhancement methods could improve the model's performance. The goal of incorporating augmented data was to assess the transformer ability to adapt to more diverse datasets. The model was trained with standard cross entropy loss to optimize its network weights during this process. The second experiment focused on analysing the effect of network depth by varying the number of encoder layers. This allowed us to examine how the depth of the ViT influences its ability to extract meaningful features and improve classification accuracy. We evaluated the effect of image size on the model's performance by training it with different input sizes which allowed us to identify the ideal trade-off between computational complexity and accuracy.

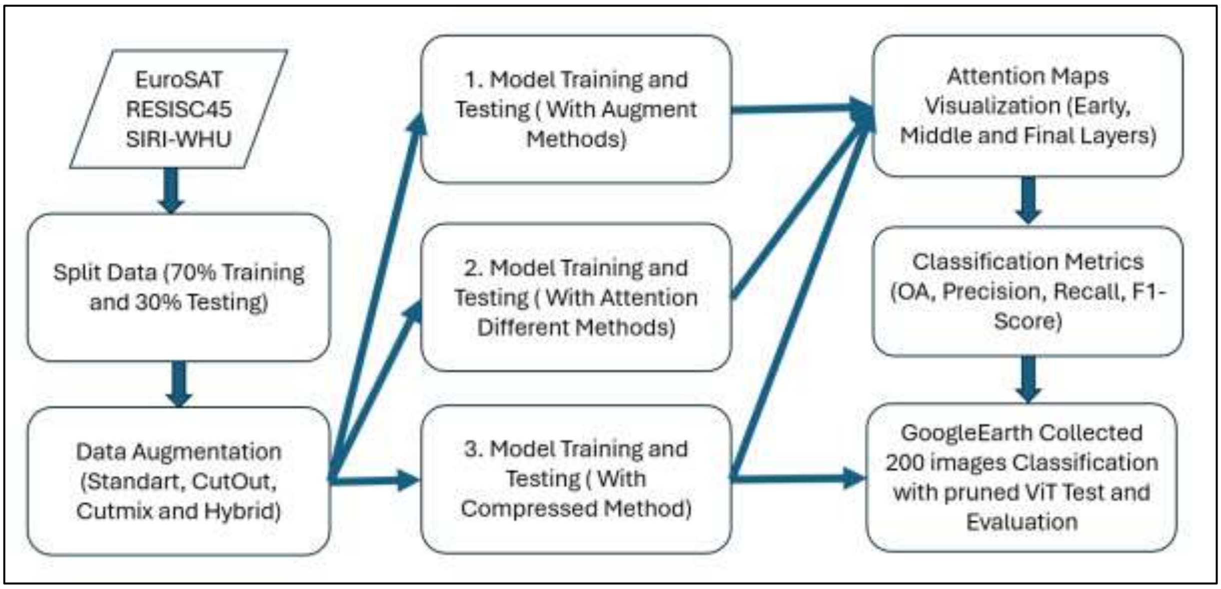
EuroSAT
RESISC45
SIRI-WHU
Split Data (70% Training and 30% Testing)
Data Augmentation (Standart, CutOut, Cutmix and Hybrid)
1. Model Training and Testing (With Augment Methods)
2. Model Training and Testing (With Attention Different Methods)
3. Model Training and Testing (With Compressed Method)
Attention Maps Visualization (Early, Middle and Final Layers)
Classification Metrics (OA, Precision, Recall, F1-Score)
GoogleEarth Collected 200 images Classification with pruned ViT Test and Evaluation

Figure 6- Agricultural Image Classification Workflow

To assess the ViT's performance, we compared its outcomes to several prominent methods in the field. This comparison yielded valuable information about its capabilities and weaknesses effectively demonstrating its ability to excel in classification tasks. These experiments collectively provided a thorough understanding of the factors influencing the accuracy and efficiency of the Vision Transformer. Our experiments utilized the ViT-Base model configured as described in the provided reference. The proposed model architecture comprises 12 encoder layers, each of which incorporates 12 separate attention heads. The embedding dimension is specified as 768 and the feed-forward subnetwork comprises 3072 units.

The model was trained on each dataset for 30 epochs, using a minibatch size of exactly 32 samples per batch. We utilized the Adam optimizer with specific settings (learning rate = 0.0001, β_1 = 0.9, β_2 = 0.999, ε = 1e-8) to ensure stable and effective training convergence. Input images were resized and constrained to a fixed dimension of 224×224 pixels and divided into 16×16 patches resulting in sequences consisting of 196 tokens for processing by the ViT architecture. Training was conducted leveraging transfer learning with initial weights pre-trained on ImageNet allowing the model to effectively adapt to agricultural imagery tasks. To systematically evaluate the model's performance employed multiple standard metrics including overall accuracy (OA), precision, recall, F1-score and confusion matrices providing a more comprehensive assessment, especially important for addressing potential class imbalance in the datasets.

Table 3- Detailed dataset information

Dataset	*Images*	*Dimensions*	*Class Names*	*Features*
EuroSAT	27000	64x64 px	Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, Sea Lake	Sentinel-2 satellite moderate spatial resolution
RESISC45	31500	256x256 px	Airport, Beach, Bridge, Chaparral, Church, Circular Farmland, Cloud, Commercial Area, Dense Residential, Desert, Forest, Freeway, Golf Course, Ground Track Field, Harbor, Island, Lake, Meadow, Mountain, Overpass, Palace, Rectangular Farmland, River, Roundabout, Runway, Sea Ice, Snowberg, Sparse Residential, Terrace, Wetland	High diversity, globally distributed satellite imagery; significant intra class variability
SIRI-WHU	2400	200x200 px	Agriculture, Commercial, Harbor, Idle Land, Industrial, Meadow, Overpass, Park, Pond, Residential, River, Water	High resolution aerial images

All experiments and implementations were conducted using PyTorch a widely used open-source deep learning library written in Python. The training and evaluation processes were executed within Google Colab's environment leveraging the Tesla T4 GPU which provided essential computational resources ensuring reproducibility and efficient handling of complex computations. Additionally, model compression was specifically performed by removing every other encoder 50% of layers effectively reducing complexity and computational overhead. This straightforward yet effective pruning strategy maintained model performance while significantly decreasing the required computational resources. Furthermore, a clear comparative summary of datasets used EuroSAT, NWPU-RESISC45 and SIRI-WHU is provided in Table 3.

3.1. Data augmentation analyses results

We performed an initial examination of the ViT by employing a low resource training setup which utilised a smaller dataset to evaluate its baseline capabilities. We chose 50% of the samples from each dataset to serve as training subsets mirroring half of the total size. This method enabled us to evaluate the model's performance with restricted data.

The network underwent a training process of 30 iterations simultaneously working with both the original and the augmented images. The impact of various data augmentation techniques on classification accuracy was assessed with specifics shown in Table 4. Methods for data augmentation typically employed involve rotation, flipping images horizontally and vertically, and applying random modifications to image brightness and colour. The Cutout technique involved the use of eight randomly positioned holes each measuring 5 by 5 pixels to conceal portions of the image. CutMix augmentation involves combining patches from various images where the blending proportion is selected randomly from a uniform distribution spanning between 0 and 1. Lastly a hybrid augmentation strategy was implemented where one of the three techniques standard, CutMix or Cutout was randomly chosen for each training batch. From the main dataset 70% of the data is reserved for training.

This systematic approach allowed us to examine the relative effectiveness of these augmentation strategies in improving model performance highlighting their role in enhancing the robustness and generalization capabilities of the ViT under constrained data scenarios.

Table 4- ViT-Base model’s classification results on different augmentation methods

Dataset	*Without Augm.*	*Standard Augm.*	*Hybrid Augm.*
NWPU	0.9260	0.9420	0.9680
SIRI-WHU	0.9350	0.9610	0.9920
EuroSat	0.9460	0.9520	0.9850

For the NWPU dataset, the baseline accuracy achieved without any data augmentation is 92.6%. Applying standard augmentation techniques such as flips, rotations and rescaling enhances the model’s performance to 94.2% indicating the value of introducing variability into the training data. However the most significant improvement is observed with hybrid augmentation, which incorporates advanced methods like synthetic sample generation or colour transformations raising the accuracy to 96.8%. This illustrates the success of hybrid augmentation in overcoming the dataset built-in difficulties including considerable diversity within each class and strong similarity between different classes.

The SIRI-WHU dataset achieves a baseline accuracy of 93.5% without any data augmentation, which is slightly higher than NWPU due to its smaller size and less complex class diversity. Standard augmentation boosts accuracy to 96.1% showing that even basic transformations help the model generalize better. With hybrid augmentation, the accuracy surges to an impressive 99.2% almost perfect classification. This substantial improvement highlights how hybrid techniques effectively overcome the limitations of a relatively smaller dataset by enriching it with diverse and varied samples.

The EuroSAT dataset starts with the highest baseline accuracy of 94.6% reflecting its well-defined and less ambiguous class structures. Standard augmentation provides a slight improvement to 95.2% suggesting that basic transformations add limited value to an already strong dataset. However, hybrid augmentation substantially enhances accuracy to 98.5% showcasing its ability to further refine the model’s generalization capabilities. Despite the dataset's strong baseline, the advanced hybrid methods help address subtler variations enabling even higher classification performance.

The table highlights several key findings relating to the impact of data augmentation on classification accuracy rates. Across all datasets augmentation consistently enhanced accuracy, underscoring its critical role in improving the model's robustness and generalization capabilities. Notably, the hybrid augmentation approach emerged as the most effective strategy, outperforming both no augmentation and standard methods by combining advanced techniques like CutMix and Cutout with traditional augmentations. Datasets with lower baseline performance such as SIRI-WHU showed the greatest gains from hybrid augmentation demonstrating its effectiveness in tackling more challenging scenarios. Overall hybrid augmentation enabled the Vision Transformer to overcome the constraints of smaller datasets and limited diversity achieving exceptional classification results in many cases.

3.2. Network attention analyses results

We investigated the output representations to gain a better comprehension of the network's behaviour and the distinct areas emphasized by each attention head across different layers through visualisation of the per layer attention maps. These visualizations were generated for the SIRI-WHU and NWPU datasets as illustrated in Figures 7 and 8, respectively. This approach allowed us to observe how the ViT progressively refines its focus on class relevant regions across the layers shedding light on the inner workings of its attention mechanism.

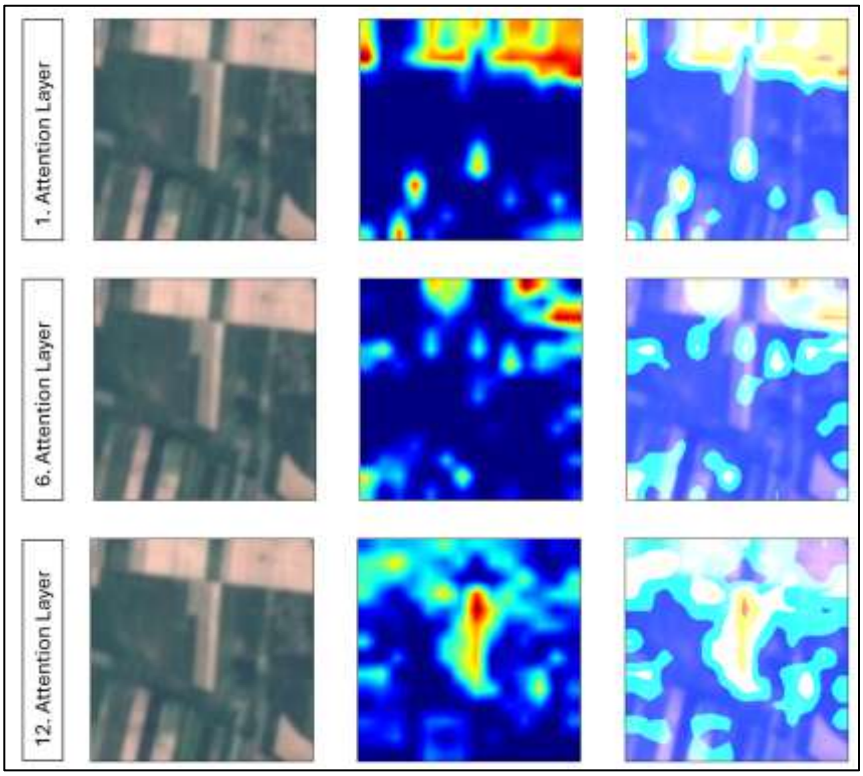

Figure 7- Different encoder layers class attention maps for agricultural class

Figure 7 showcases the progressive refinement of attention maps across different layers of a model applied to an agricultural field image highlighting its hierarchical learning process. In the initial layer, the attention is broadly distributed across the image capturing general textures, patterns and colours without distinguishing relevant from irrelevant regions. By the middle layer, the focus becomes more structured concentrating on distinct field features such as crop rows, boundaries or planting patterns while gradually disregarding background noise. In the final layer, the attention is sharply localized on critical and class representative regions such as prominent crop textures or field edges effectively isolating the features essential for classification. This transition from dispersed to highly targeted focus underscores the model's ability to identify and prioritize meaningful features for precise agricultural analysis.

Figure 8 shows classification analysis of the agricultural field images the attention maps clearly illustrate the progressive refinement of the model’s focus throughout its layers. In the initial layers, the attention is broadly distributed across the image capturing general patterns, textures and colour variations, including areas unrelated to the agricultural fields. At this stage the model appears to focus on global image features reflecting a generalized pattern recognition process. In the middle layers, such as layers 5–7 the attention gradually shifts towards more specific and distinctive characteristics of the agricultural fields such as planting rows, boundary lines or structural elements within the field. Irrelevant background areas, such as surrounding vegetation or soil textures are increasingly ignored. Allowing the model to sharpen its focus on features that are more representative of the field’s unique patterns.

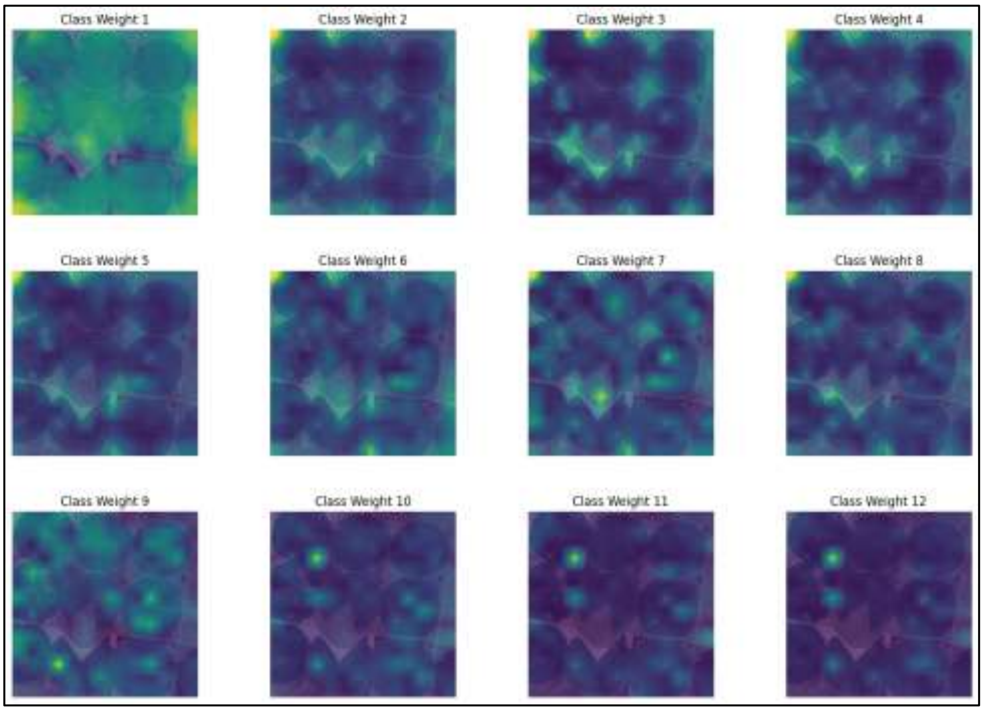

Figure 8- Agricultural image class attention maps result

By the final layers, the attention maps become highly concentrated on the most critical and class representative regions of the agricultural fields isolating distinctive features such as crop textures planting rows or other identifiable field structures. This progression from generalized to highly specific attention highlights the hierarchical learning capability of the model and its ability to identify and prioritize key features essential for accurate classification. These insights demonstrate the model's effectiveness in learning complex spatial patterns and its potential for robust agricultural field analysis.

3.3. Different image size results

Our investigation into the ViT's performance involved conducting tests with two specific image resolutions, 224 by 224 pixels and 384 by 384 pixels. Since the model was pretrained on ImageNet with 384×384 images this provided a baseline for evaluating how training with smaller images affects accuracy. The results, summarized in Table 5 reveal that increasing the image size consistently improved accuracy across all datasets. Using higher image resolutions continues to enhance classification accuracy across the datasets NWPU, SIRI-WHU and EuroSAT while also increasing computational time though the impact varies across datasets. For NWPU-RESISC45, the accuracy rises from 96.8% at 224x224 resolution to 98.8% at 384x384 resolution highlighting the model's ability to better distinguish complex features with higher spatial detail. However, this improvement nearly doubles the training time increasing from 1 hour 10 minutes to 2 hours 20 minutes. Similarly, SIRI-WHU shows an improvement from 99.2% to 99.7% with the resolution increase but training time grows from 1 hour 23 minutes to 2 hours 33 minutes. These results indicate that while higher resolutions boost accuracy they impose notable computational costs, particularly for datasets with diverse and nuanced class structures.

For the EuroSAT dataset, which has a simpler structure and fewer image variations, the accuracy improves from 98.5% at 224x224 resolution to 99.5% at 384x384. Despite this increase the training times remain relatively short compared to the other datasets rising from 24 minutes to 40 minutes. This reflects EuroSAT's smaller dataset size and less complex image features making it computationally efficient even with higher resolutions. These findings emphasize the trade-off between accuracy and computational efficiency while larger image resolutions significantly enhance the performance of datasets with complex spatial details the corresponding increase in training time must be carefully considered. For resource constrained applications balancing resolution with computational efficiency is critical particularly for datasets like NWPU-RESISC45 where the gains are substantial and SIRI-WHU where high baseline performance sees diminishing returns.

Table 5- Training accuracy and times different image size

Dataset	*224x224*	*384x384*
NWPU	0.9680 (01:10:38)	0.9880 (02:20:04)
SIRI-WHU	0.9920 (01:23:57)	0.9970 (02:33:40)
EuroSat	0.9850 (24:17)	0.9950 (40:12)

The higher resolution provides richer details for the model to analyse enabling it to capture more discriminative features and enhance classification accuracy. Yet the added computational load highlights a trade-off between achieving better performance and managing training time. Deciding on the image size involves balancing the need for precision against available computational resources making it a key consideration in practical applications.

3.4. Compressed network classification results

To evaluate the effectiveness of model compression techniques in agricultural image classification, a pruned version of the ViT-Base-V16 model configured at an input resolution of 224×224 pixels was tested across three benchmark datasets. The pruning strategy involved the systematic removal of 50% of the encoder layers reducing model complexity while preserving core attention mechanisms. Despite the significant reduction in network depth the pruned model achieved remarkably high classification accuracy: 95.66% on NWPU-RESISC45, 97.90% on SIRI-WHU and 95.60% on EuroSAT. These results demonstrate that aggressive pruning can be implemented without severely compromising performance particularly when the model is pretrained on large-scale datasets such as ImageNet. The minimal drop in accuracy especially on the more structured EuroSAT and SIRI-WHU datasets suggests that the compressed model retains its ability to extract essential spatial features relevant for land use and land cover classification.

Table 6- Comparison of the proposed model accuracy with State-Of-The-Art (SOTA) techniques

Model	*NWPU*	*SIRI-WHU*	*EuroSat*
Vgg16 (Tan & Le 2019)	0.8740	0.8430	0.8160 (Sandler et al. 2018)
InceptionV3 (Szegedy et al. 2016)	0.8900	0.8640	0.8630
RexNet (Tan & Le 2019)	0.9160	0.9040	0.9110
ResNet50 (Thirumaladevi et al. 2023)	0.9140 (Tang et al. 2018)	0.9010	0.8810
EfficientNet-B0 (Tan & Le 2019)	0.9410	0.9350 (He et al. 2016)	0.9060 (Zhang et al. 2018)
Deit-Tiny (Martins et al. 2020)	0.9430	0.9250	0.9580
ViT-Base-16 (384x384)	0.9880	0.9970	0.9950
ViT-Base-32 (224x224)	0.9680	0.9920	0.9850
ViT-Base-V16 (224x224) %50 Pruned	**0.9566**	**0.9790**	**0.9560**

The primary novelty of this study lies in the integration of three complementary strategies to enhance the efficiency and generalization of ViT architectures for agricultural image classification. First, the application of ViT based models particularly ViT-based on remote sensing imagery for agricultural land use and land cover classification represents a relatively unexplored area expanding the use of transformer models beyond traditional natural image domains. Second, the research demonstrates that a structured pruning strategy involving the removal of 50% of the encoder layers can significantly reduce computational complexity without incurring substantial performance degradation. This finding is particularly important for enabling the deployment of ViT models in real world scenarios where computational resources are limited. Third the study introduces and validates a hybrid data augmentation pipeline combining standard techniques with CutMix and Cutout. This combination proves to be highly effective in improving classification accuracy especially under constrained training data scenarios. The synergy of these three contributions ViT adaptation efficient pruning and hybrid augmentation constitutes the core innovation of the research offering a scalable and practical solution for precision agriculture applications using remote sensing imagery.

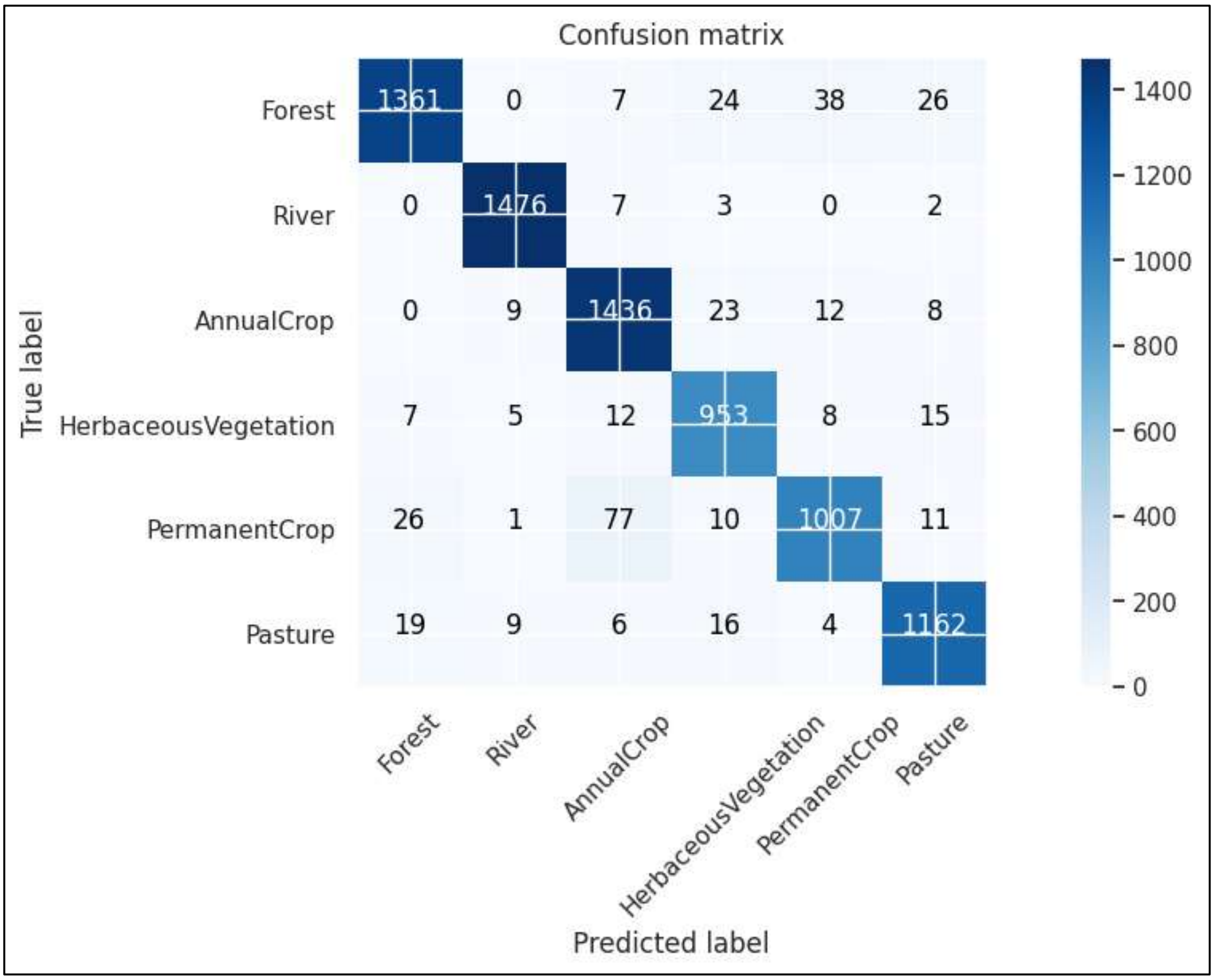

Figure 9- Confusion matrix tested 7780 images from EuroSAT with pruned ViT

The confusion matrix generated from the evaluation of 7780 test images from the EuroSAT dataset demonstrates the strong classification capability of the pruned ViT-Base-V16 model across six major land cover classes forest, river, annual crop, herbaceous vegetation, permanent crop and pasture. The model achieved precision of 95.07%, recall of 95.05% and F1-score of 95.04%, indicating a highly balanced and robust classification performance. Among the classes river achieved the highest classification accuracy with 1476 correctly identified samples and minimal misclassification reflecting the distinct spectral characteristics of water bodies in remote sensing imagery. Forest and pasture classes also demonstrated strong performance with 1361 and 1162 correctly classified samples respectively although minor confusion was observed between classes while generally classified with high accuracy 1436 and 953 correct predictions respectively showed moderate confusion with each other consistent with the spectral and structural similarities typical of vegetation rich land cover types. Figure 9 shows confusion matrix for all classes.

Permanent crop classification presented slightly higher confusion particularly with annual crop and forest classes likely due to overlapping spectral signatures during certain agricultural seasons. Nevertheless, the overall high concentration of true positives along the diagonal of the matrix indicates the model’s strong ability to distinguish between most land cover categories with high confidence. These results confirm that the pruned ViT-Base-V16 model maintains its classification strength even after substantial compression offering a scalable and efficient solution for real world agricultural monitoring and land management applications. Minor misclassification patterns suggest that future improvements could be achieved through the integration of multispectral or temporal information particularly for vegetation classes with subtle intra class variability.

Table 7- Pruned ViT tested on EuroSAT dataset metrics

Class	*Precision*	*Recall*	*F1-Score*	*Support*
Forest	0.96	0.93	0.95	1456
River	0.98	0.99	0.99	1488
Annual Crop	0.93	0.97	0.95	1488
Herbaceous Vegetation	0.93	0.95	0.94	1000
Permanent Crop	0.94	0.89	0.92	1132
Pasture	0.95	0.96	0.95	1216
Overall Accuracy	-	-	**0.95**	7780
Macro Average	0.95	0.95	0.95	7780
Weighted Average	0.95	0.95	0.95	7780

In table 7, pruned ViT-Base-V16 model achieved an overall accuracy of 95% with macro averaged precision, recall and F1-scores of 0.95 demonstrating consistent performance across all classes. The River class showed the highest results precision 98% and recall 99% highlighting the model’s strong ability to distinguish water bodies. Forest and pasture were also classified with high precision and recall 95% while annual crop exhibited a high recall 97% but slightly lower precision 93% indicating occasional confusion with similar vegetation types. Herbaceous vegetation maintained balanced precision and recall 94% whereas permanent crop had a lower recall 89% reflecting challenges in distinguishing permanent plantations. Overall, the results validate the pruned ViT-Base-V16 model as an accurate and computationally efficient solution for agricultural land cover classification with only minor confusion among spectrally similar vegetation classes.

3.5. Newly created dataset classification results

To assess the generalization capability of the pruned ViT model a new dataset was compiled consisting of 200 agricultural area images collected from Google Earth. Each image resized to a fixed dimension of 224×224 pixels. The pruned ViT model optimized previously by removing 50% of its encoder layers to reduce computational complexity was evaluated on this independent dataset. Remarkably the model accurately classified 198 out of the 200 images achieving an impressive 99% overall accuracy. This performance underscores the model’s robustness and exceptional generalization ability beyond standard benchmark datasets. Furthermore, it demonstrates the practical applicability of the pruned ViT model to real world remote sensing scenarios validating its effectiveness for agricultural monitoring and precision agriculture applications. Figure 10 shows predicted real world images.

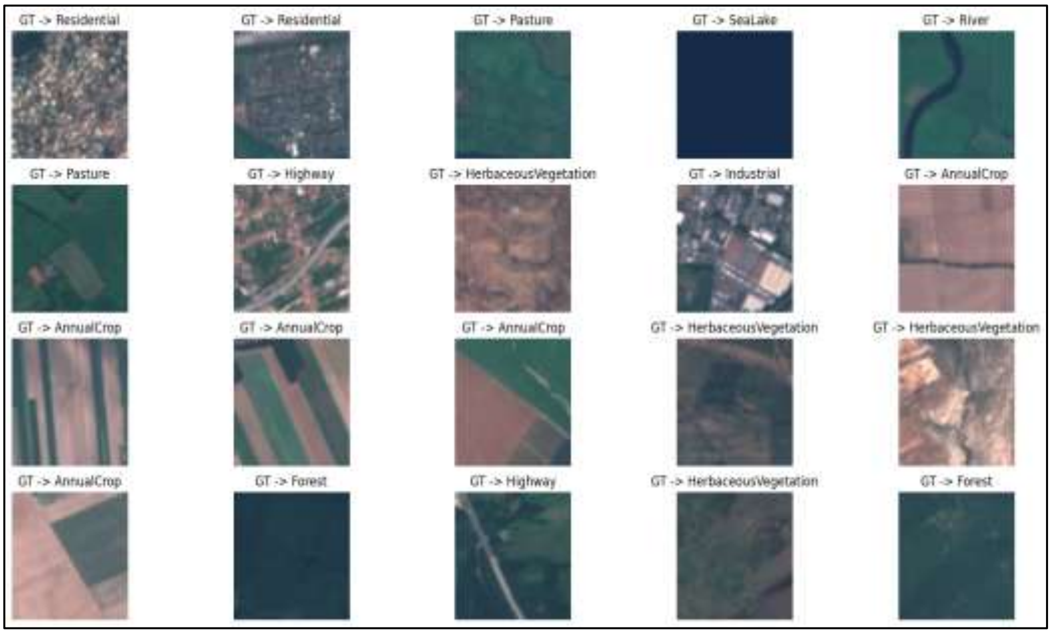

Figure 10- Generalisation dataset classification result images

4. Discussion

The results of this study not only confirm the superior classification performance of the pruned ViT model but also provide key insights into the broader implications of model optimization strategies in remote sensing. This research introduces several novel contributions. First the adaptation of ViT architectures originally designed for natural image recognition to agricultural remote sensing represents a relatively unexplored domain. The ability of ViT to capture global spatial relationships enables effective representation of complex landscape patterns particularly in large-scale satellite imagery. Second the implementation of a structured pruning approach in which 50% of the encoder layers were removed substantially reduced computational demands with only a marginal decline in accuracy. This makes the model highly suitable for deployment in real-world systems with constrained computational resources. Third the study introduces a hybrid data augmentation pipeline combining traditional transformations with CutMix and Cutout which demonstrated strong efficacy under limited data conditions by improving generalization and robustness.

The comparative analysis across the NWPU-RESISC45, SIRI-WHU and EuroSAT datasets reveals significant performance differences influenced by both model architecture and dataset complexity. Transformer based models consistently outperformed CNNs confirming their strength in modelling long range dependencies and handling spatial heterogeneity. For example, ViT-Base-16 (384×384) achieved 98.8% accuracy on NWPU characterized by high intra class diversity and inter class similarity illustrating the benefit of high-resolution inputs and attention based modelling. The 224×224 variant of ViT-Base also performed well 96.8% while lightweight models like EfficientNet-B0 94.1% and DeiT-Tiny 94.3% offered competitive results showing promise for resource efficient applications. On the SIRI-WHU dataset which contains fewer classes and simpler spatial structures ViT-Base-16 achieved near perfect performance 99.7% demonstrating its capability to capture fine grained details. The pruned ViT-Base-V16 (224×224) also performed strongly 97.9% supporting the effectiveness of compression for efficient inference. CNN-based models such as EfficientNet-B0 93.5% and DeiT-Tiny 92.5% remained competitive while traditional models like VGG16 84.3% and InceptionV3 86.4% lagged behind underscoring their limited ability to model subtle spatial patterns. Similarly, on the EuroSAT dataset which features well separated land cover classes and lower structural complexity. ViT-Base-16 once again achieved the highest accuracy 99.5%. DeiT-Tiny 95.8% and ViT-Base (224×224) 98.5% followed closely. Despite the dataset's simplicity the consistent outperformance of transformer-based models reflect their robustness and generalization capacity. While EfficientNet-B0 90.6% and RexNet 91.1% yielded satisfactory performance older models such as VGG16 81.6% and InceptionV3 86.3% performed considerably worse aligning with their architectural limitations in capturing hierarchical and contextual relationships.

A closer examination of model behaviour across datasets highlights that differences in performance are not solely architectural but are also shaped by the inherent properties of the datasets. The higher performance of ViT-based on EuroSAT compared to NWPU. Attributed to EuroSAT's lower class complexity, well-defined semantic boundaries and consistent spectral characteristics. In contrast NWPU contains more complex and globally distributed scenes with overlapping class semantics, increasing classification difficulty. These observations underscore the importance of aligning model complexity and capacity with dataset characteristics when deploying deep learning models for remote sensing tasks.

Overall, the findings emphasize that transformer-based models particularly at higher resolutions currently represent the state-of-the-art in remote sensing image classification. However, models such as EfficientNet-B0 and DeiT-Tiny provide viable alternatives for real-time or mobile applications. Furthermore, the success of the pruned ViT-Base-V16 in maintaining accuracy while halving model size illustrates the viability of compression techniques for operational use. This analysis supports the conclusion that model selection should be guided by both dataset complexity and application specific resource constraints, particularly in precision agriculture and other geospatial intelligence domains.

5. Conclusions

This study presents a comprehensive evaluation of deep learning models particularly ViTs for land use and land cover classification using remote sensing imagery. Experiments were conducted on three publicly available datasets NWPU-RESISC45, SIRI-WHU and EuroSAT demonstrating that ViT-Base-16 (384×384) consistently achieves the highest classification accuracy across all datasets. Its success can be attributed to its strong capacity to model global spatial dependencies making it especially effective on complex datasets with high intra-class variability such as NWPU. The study also highlights the practical value of lightweight alternatives such as EfficientNet-B0 and DeiT-Tiny. These models deliver competitive performance while maintaining low computational demands rendering them ideal for real world deployment scenarios including edge devices, mobile systems and remote sensing operations in resource limited environments.

A key contribution of this work lies in the successful implementation of a 50% pruning strategy on the ViT-Based model which significantly reduces model complexity and computational overhead while maintaining high classification accuracy 95.66% on NWPU, 97.90% on SIRI-WHU and 95.60% on EuroSAT. This makes the compressed model particularly suitable for use in real time agricultural monitoring applications especially in areas where access to high performance hardware is limited. Importantly the model was also tested on a custom Google Earth-based agricultural image dataset achieving a 99% accuracy on 200 previously unseen images. This confirms the model's strong generalization ability and its applicability to real-world remote sensing imagery beyond benchmark datasets. In addition, a hybrid data augmentation pipeline integrating CutMix, Cutout and standard techniques proved especially effective in improving model robustness and generalization particularly under limited data scenarios. This further enhances the model's suitability for use in operational agricultural systems where labelled data may be scarce or expensive to obtain.

In conclusion, the study introduces a practical and scalable transformer based solution for precision agriculture capable of operating efficiently in low resource settings while maintaining state-of-the-art performance. Future research could focus on integrating multispectral or hyperspectral data into transformer based models to exploit rich spectral features in agricultural imagery. The use of temporal data may further support time-series analysis for crop monitoring and yield forecasting. Additionally, implementing quantization and knowledge distillation can enhance model deployability on real time and edge devices. The application of active learning strategies may improve labeling efficiency in data scarce environments. Finally, deploying the pruned ViT models in real world agricultural systems, such as UAV based crop monitoring platforms or remote sensing decision support tools would offer valuable insight into their operational effectiveness and scalability.

Data statement: The data that support the findings of this study are openly available in [EuroSAT] at [https://github.com/phelber/EuroSAT], [NWPU-RESISC45] at [https://arxiv.org/abs/1703.00121], [SIRI-WHU] at [http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/Num/Google.html].

References

Ahmad M, Shabbir S, Roy S K, Hong D, Wu X, Yao J & Chanussot J (2022). Hyperspectral Image Classification - Traditional to Deep Models: A Survey for Future Prospects. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15: 968–999. doi: 10.1109/JSTARS.2021.3133021

Albarakati H M, Khan M A, Hamza A, Khan F, Kraiem N, Jamel L & Alroobaea R (2024). A Novel Deep Learning Architecture for Agriculture Land Cover and Land Use Classification from Remote Sensing Images Based on Network-Level Fusion of Self-Attention Architecture. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 6338–6353. doi: 10.1109/JSTARS.2024.3369950

Albert A, Kaur J & Gonzalez M C (2017). Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Part F129685, 1357–1366. Association for Computing Machinery. doi: 10.1145/3097983.3098070

Alzahrani M S & Alsaade F W (2023). Transform and Deep Learning Algorithms for the Early Detection and Recognition of Tomato Leaf Disease. Agronomy 13(5). doi: 10.3390/agronomy13051184

Attri I, Awasthi L K, Sharma T P & Rathee P (2023). A review of deep learning techniques used in agriculture. Ecological Informatics, Vol. 77. Elsevier B.V. doi: 10.1016/j.ecoinf.2023.102217

Basu S, Ganguly S, Mukhopadhyay S, DiBiano R, Karki M & Nemani R (2015). DeepSat - A learning framework for satellite imagery. GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, 03-06-November-2015. Association for Computing Machinery. doi: 10.1145/2820783.2820816

Bazi Y, Bashmal L, Al Rahhal M M, Dayil R Al & Ajlan N Al (2021). Vision transformers for remote sensing image classification. Remote Sensing, 13(3): 1–20. doi: 10.3390/rs13030516

Beyer L, Zhai X & Kolesnikov A (2022). Better plain ViT baselines for ImageNet-1k. Retrieved from http://arxiv.org/abs/2205.01580

Bowles C, Chen L, Guerrero R, Bentley P, Gunn R, Hammers A & Rueckert D (2018). GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks. Retrieved from http://arxiv.org/abs/1810.10863

Celik F, Balik Sanli F & Bozic D (2024). Transformer Networks to Classify Weeds and Crops in High-Resolution Aerial Images From North-East Serbia. Turkish Journal of Field Crops 29(2): 112–120. doi: 10.17557/tjfc.1511404

Chen W, Du X, Yang F, Beyer L, Zhai X, Lin T.-Y.& Zhou D (2021). A Simple Single-Scale Vision Transformer for Object Localization and Instance Segmentation. Retrieved from http://arxiv.org/abs/2112.09747

Cheng G, Han J & Lu X (2017, October 1). Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proceedings of the IEEE, Vol. 105, pp. 1865–1883. Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/JPROC.2017.2675998

Cubuk E D, Zoph B, Mane D, Vasudevan V & Le Q V (2019). Autoaugment: Learning augmentation strategies from data. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June, 113–123. IEEE Computer Society. doi: 10.1109/CVPR.2019.00020

DeVries T & Taylor G W (2017). Improved Regularization of Convolutional Neural Networks with Cutout. Retrieved from http://arxiv.org/abs/1708.04552

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T & Houlsby N (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Retrieved from http://arxiv.org/abs/2010.11929

Gandhar A, Gupta K, Pandey A K & Raj D (2024, June 1). Fraud Detection Using Machine Learning and Deep Learning. SN Computer Science, Vol. 5. Springer. doi: 10.1007/s42979-024-02772-x

Gong B, Dai K, Shao J, Jing L & Chen Y (2023). Fish-TViT: A novel fish species classification method in multi water areas based on transfer learning and vision transformer. Heliyon 9(6). doi: 10.1016/j.heliyon.2023.e16761

Guo N, Jiang M, Gao L, Li K, Zheng F, Chen X & Wang M (2023). HFCC-Net: A Dual-Branch Hybrid Framework of CNN and CapsNet for Land-Use Scene Classification. Remote Sensing, 15(20). doi: 10.3390/rs15205044

Hamza A, Khan M A, Ur Rehman S, Albarakati H M, Alroobaea R, Baqasah A M & Masood A (2023). An Integrated Parallel Inner Deep Learning Models Information Fusion with Bayesian Optimization for Land Scene Classification in Satellite Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 9888–9903. doi: 10.1109/JSTARS.2023.3324494

Hamza A, Khan M A, Ur Rehman S, Al-Khalidi M, Alzahrani A I, Alalwan N & Masood A (2024). A Novel Bottleneck Residual and Self-Attention Fusion-Assisted Architecture for Land Use Recognition in Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17: 2995–3009. doi: 10.1109/JSTARS.2023.3348874

Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z & Tao D (2020). A Survey on Visual Transformer. doi: 10.1109/TPAMI.2022.3152247

He K, Zhang X, Ren S & Sun J (2016a). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770–778. IEEE Computer Society. doi: 10.1109/CVPR.2016.90

He K, Zhang X, Ren S & Sun J (2016b). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770–778. IEEE Computer Society. doi: 10.1109/CVPR.2016.90

Helber P, Bischke B, Dengel A & Borth D (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(7): 2217–2226. doi: 10.1109/JSTARS.2019.2918242

Hinton G, Vinyals O & Dean J (2015). Distilling the Knowledge in a Neural Network. Retrieved from http://arxiv.org/abs/1503.02531

Huete A, Didan K, Miura T, Rodriguez E P, Gao X & Ferreira L G (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Retrieved from www.elsevier.com/locate/rse

Jackson P T, Atapour-Abarghouei A, Bonner S, Breckon T & Obara B (2018). Style Augmentation: Data Augmentation via Style Randomization. Retrieved from http://arxiv.org/abs/1809.05375

Joshi A, Pradhan B, Gite S & Chakraborty S (2023, April 1). Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review. Remote Sensing, Vol. 15. MDPI. doi: 10.3390/rs15082014

Khan S D & Basalamah S (2023). Multi-Branch Deep Learning Framework for Land Scene Classification in Satellite Imagery. Remote Sensing, 15(13). doi: 10.3390/rs15133408

Khan S, Naseer M, Hayat M, Zamir S W, Khan F S & Shah M (2021). Transformers in Vision: A Survey. doi: 10.1145/3505244

Li X & Li S (2022). Transformer Help CNN See Better: A Lightweight Hybrid Apple Disease Identification Model Based on Transformers. Agriculture (Switzerland), 12(6). doi: 10.3390/agriculture12060884

Martins V S, Kaleita A L, Gelder B K, da Silveira H L F & Abe C A (2020). Exploring multiscale object-based convolutional neural network (multi-OCNN) for remote sensing image classification at high spatial resolution. *ISPRS Journal of Photogrammetry and Remote Sensing*, 168: 56–73. doi: 10.1016/j.isprsjprs.2020.08.004

Maurício J, Domingues I & Bernardino J (2023, May 1). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Applied Sciences (Switzerland), Vol. 13. MDPI. doi: 10.3390/app13095521

Miotto R, Wang F, Wang S, Jiang X & Dudley J T (2017). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6): 1236–1246. doi: 10.1093/bib/bbx044

Nguyen T T, Hoang T D, Pham M T, Vu T T, Nguyen T H, Huynh Q T & Jo J (2020). Monitoring agriculture areas with satellite images and deep learning. Applied Soft Computing Journal, 95. doi: 10.1016/j.asoc.2020.106565

Nie J, Yuan Y, Li Y, Wang H, Li J, Wang Y & Ercisli S (2024). Few-shot Learning in Intelligent Agriculture: A Review of Methods and Applications. *Tarim Bilimleri Dergisi* 30(2): 216–228. doi: 10.15832/ankutbd.1339516

Papageorgiou E I, Markinos A T & Gemtos T A (2011). Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application. Applied Soft Computing Journal, 11(4): 3643–3657. doi: 10.1016/j.asoc.2011.01.036

Park S, Im J, Park S, Yoo C, Han H & Rhee J (2018). Classification and mapping of paddy rice by combining Landsat and SAR time series data. Remote Sensing, 10(3). doi: 10.3390/rs10030447

Ranftl R, Bochkovskiy A & Koltun V (2021). Vision Transformers for Dense Prediction. Retrieved from https://github.com/intel-isl/DPT.

Reedha R, Dericquebourg E, Canals R & Hafiane A (2022). Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images. Remote Sensing, 14(3). doi: 10.3390/rs14030592

Sandler M, Howard A, Zhu M, Zhmoginov A & Chen L C (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 4510–4520. IEEE Computer Society. doi: 10.1109/CVPR.2018.00474

Shastry K A, Sanjay H A & Deexith G (2017). Quadratic-radial-basis-function-kernel for classifying multi-class agricultural datasets with continuous attributes. Applied Soft Computing Journal, 58: 65–74. doi: 10.1016/j.asoc.2017.04.049

Shin H, Jeon S, Seol Y, Kim S & Kang D (2023). Vision Transformer Approach for Classification of Alzheimer’s Disease Using 18F-Florbetaben Brain Images. Applied Sciences (Switzerland), 13(6). doi: 10.3390/app13063453

Singh G, Singh S, Sethi G & Sood V (2022). Deep Learning in the Mapping of Agricultural Land Use Using Sentinel-2 Satellite Data. Geographies, 2(4): 691–700. doi: 10.3390/geographies2040042

Slavkovikj V, Verstockt S, De Neve W, Van Hoecke S & Van De Walle R (2015). Hyperspectral image classification with convolutional neural networks. MM 2015 - Proceedings of the 2015 ACM Multimedia Conference, 1159–1162. Association for Computing Machinery, Inc. doi: 10.1145/2733373.2806306

Suh H K, IJsselmuiden J, Hofstee J W & van Henten E J (2018). Transfer learning for the classification of sugar beet and volunteer potato under field conditions. Biosystems Engineering, 174: 50–65. doi: 10.1016/j.biosystemseng.2018.06.017

Suravarapu V K & Patil H Y (2023). Person Identification and Gender Classification Based on Vision Transformers for Periocular Images. Applied Sciences (Switzerland), 13(5). doi: 10.3390/app13053116

Szegedy C, Vanhoucke V, Ioffe S, Shlens J & Wojna Z (2016). Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2818–2826. IEEE Computer Society. doi: 10.1109/CVPR.2016.308

Tan M & Le Q V (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Retrieved from http://arxiv.org/abs/1905.11946

Tang X, Zhang X, Liu F & Jiao L (2018). Unsupervised deep feature learning for remote sensing image retrieval. Remote Sensing, 10(8). doi: 10.3390/rs10081243

Temenos A, Temenos N, Kaselimi M, Doulamis A & Doulamis N (2023). Interpretable Deep Learning Framework for Land Use and Land Cover Classification in Remote Sensing Using SHAP. IEEE Geoscience and Remote Sensing Letters, 20. doi: 10.1109/LGRS.2023.3251652

Thakur P S, Chaturvedi S, Khanna P, Sheorey T & Ojha A (2023). Vision transformer meets convolutional neural network for plant disease classification. Ecological Informatics, 77. doi: 10.1016/j.ecoinf.2023.102245

Thirumaladevi S, Veera Swamy K & Sailaja M (2023a). Remote sensing image scene classification by transfer learning to augment the accuracy. Measurement: Sensors, 25. doi: 10.1016/j.measen.2022.100645

Thirumaladevi S, Veera Swamy K & Sailaja M (2023b). Remote sensing image scene classification by transfer learning to augment the accuracy. Measurement: Sensors, 25. doi: 10.1016/j.measen.2022.100645

Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L & Polosukhin I (2017). Attention Is All You Need.

Vohra R & Tiwari K C (2023). Land cover classification using multi-fusion based dense transpose convolution in fully convolutional network with feature alignment for remote sensing images. Earth Science Informatics, 16(1): 983–1003. doi: 10.1007/s12145-022-00891-8

Xia Z, Pan X, Song S, Erran Li, L & Huang G (2022). Vision Transformer with Deformable Attention. Retrieved from https://github.com/LeapLabTHU/DAT.

Yun S, Han D, Oh S J, Chun S, Choe J & Yoo Y (2019). CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. Retrieved from http://arxiv.org/abs/1905.04899

Zahra U, Khan M A, Alhaisoni M, Alasiry A, Marzougui M & Masood A (2024). An Integrated Framework of Two-Stream Deep Learning Models Optimal Information Fusion for Fruits Disease Recognition. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 3038–3052. doi: 10.1109/JSTARS.2023.3339297

Zhai X, Kolesnikov A, Houlsby N & Beyer L (2021). Scaling Vision Transformers. Retrieved from http://arxiv.org/abs/2106.04560

Zhang H, Cisse M, Dauphin Y N & Lopez-Paz D (2017). mixup: Beyond Empirical Risk Minimization. Retrieved from http://arxiv.org/abs/1710.09412

Zhang M, Lin H, Wang G, Sun H & Fu J (2018). Mapping paddy rice using a Convolutional Neural Network (CNN) with Landsat 8 datasets in the Dongting Lake Area, China. Remote Sensing, 10(11). doi: 10.3390/rs10111840

Zhang X, Zhou X, Lin M & Sun J (2018). ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 6848–6856. IEEE Computer Society. doi: 10.1109/CVPR.2018.00716

Zhao B, Zhong Y, Xia G S & Zhang L (2016). Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 54(4), 2108–2123. doi: 10.1109/TGRS.2015.2496185

Zhao S, Tu K, Ye S, Tang H, Hu Y & Xie C (2023, November 3). Land Use and Land Cover Classification Meets Deep Learning: A Review. Sensors (Basel, Switzerland), Vol. 23. doi: 10.3390/s23218966

Zuo S, Xiao Y, Chang X & Wang X (2022). Vision transformers for dense prediction: A survey. Knowledge-Based Systems, 253. doi: 10.1016/j.knosys.2022.109552