Turkish Journal of FIELD CROPS

Turk J Field Crops
2024, 29(2), 112-120
DOI: 10.17557/tjfc.1511404

TRANSFORMER NETWORKS TO CLASSIFY WEEDS AND CROPS IN HIGH-RESOLUTION AERIAL IMAGES FROM NORTH-EAST SERBIA

Fatih CELIK[1], Fusun BALIK SANLI[1], Dragana BOZIC[2]*

[1] Yildiz Technical University, Department of Geomatic Engineering, Istanbul 34210, Turkey
[2] University of Belgrade, Faculty of Agriculture, Nemanjina 6, 11080 Belgrade, Serbia
Corresponding author: fatih.celik1@std.yildiz.edu.tr

Received: 05.07.2024

ABSTRACT

The intricate backgrounds present in crop and field images, coupled with the minimal contrast between weed-infested areas and the background, can lead to considerable ambiguity. This, in turn, poses a significant challenge to the resilience and precision of crop identification models. Identifying and mapping weeds are pivotal stages in weed control, essential for maintaining crop health. A multitude of research efforts underscore the significance of leveraging remote sensing technologies and sophisticated machine learning algorithms to enhance weed management strategies. Deep learning techniques have demonstrated impressive effectiveness in a range of agricultural remote sensing applications, including plant classification and disease detection. High-resolution imagery was collected using a UAV equipped with a high-resolution camera, which was strategically deployed over weed, sunflower, tobacco and maize fields to collect data. The VIT models achieved commendable levels of accuracy, with test accuracies of 92.97% and 90.98% in their respective evaluations. According to the experimental results, transformers not only excel in crop classification accuracy, but also achieve higher accuracy with a smaller sample size. Swin-B16 achieved an accuracy of 91.65% on both the training and test datasets. Compared to the other two ViT models, the loss value is significantly lower by half, at 0.6450.

Keywords: agriculture; drone; image classification; multi-head attention; remote sensing; vision transformers.

Abbreviations: unmanned aerial vehicles (UAV), deep learning (DL), natural language processing (NLP), vision transformers (ViT), convolutional neural networks (CNN), multilayer perceptron (MLP).

INTRODUCTION

In an ever-changing and progressive industrial landscape, agriculture plays a significant role in overcoming numerous challenges to achieve high yields while maintaining plant growth and quality standards to meet the demands of both society and the market. Yet, the age-old problem persists in modern agriculture: an over-reliance on pesticide interventions to boost production capacity, enhance quality, and combat unwanted plant growth, especially weeds (Grammatikis et al. 2020; Ustuner et al. 2020). Weeds compete with primary crops for vital development resources such as water, nutrients and sunlight. They pose a significant challenge to the outlook for agricultural production. The widespread use of herbicides in sprayed fields increases environmental damage such as air, water and soil pollution. Some weed species develop resistance to these chemicals. This continuing trend could threaten crop yields if weed resistance is fully realized. Site-specific weed and crop control management needs to be developed as an area of research (Iqbal et al. 2019; Vrbničanin et al. 2017).

One effective solution is the use of automated crop monitoring and inspection systems which offer promising environmental and economic benefits. The advantage of using robotic technology is that it reduces labour costs and minimises the use of herbicides. In addition, weeds often have similar colour, texture and shape characteristics as crops. Automated weed control systems face the challenge of identifying and mapping weeds in the field (Iqbal et al. 2019; Wu et al. 2020). Unmanned Aerial Vehicles (UAV) utilise RGB and additional multispectral imagery to map weed density in fields. UAVs capture images as they fly over fields at different altitudes (Huang et al. 2018a, 2020b, 2018c). It uses learning algorithms to distinguish and classify weeds from crops by segmenting these large images into smaller, regular frames for effective analysis (dos Santos Ferreira et al. 2017a, 2019b).

Unlike conventional machine learning methods, which heavily depend on meticulous feature engineering, deep learning (DL) techniques autonomously extract features from images, yielding a wealth of detailed information. This results in notably enhanced performance, particularly on larger and more diverse datasets. DL has emerged as a

transformative force across numerous domains, including agriculture object detection and recognition (Hasan et al. 2021; Lecun et al. 2015). Convolutional neural networks have achieved superiority and success in tasks by extracting features from images in object detection and image classification processes through convolution filters by utilizing principles such as local connectivity, weight sharing and translation equivalence (Lecun et al. 2015; He et al. 2016). In particular, convolution-based architectural networks, including frequently used models such as VGG-16, GoogLeNet, ResNet-50, ResNet-101, AlexNet and Inception-v3, have been widely used for weed detection or classification (Madsen et al. 2020; Szegedy et al. 2016; Niu et al. 2021).

Attention mechanisms, developed primarily for natural language processing (NLP), have made significant advances and have shown significant performance improvements compared to previous versions (Niu et al. 2021; Vaswani et al. 2017). However, its adaptation to vision-related tasks has limited the significant computational demands that correspond to the higher number of pixels in images compared to NLP word studies, making traditional attention models unsuitable for use (Hasan et al. 2021; Lu et al. 2020). A significant increase in the use of transformer-attention models can be seen in computer vision with the advent of the sign-relative transformer (Li et al., 2022). Unlike CNN-based methods operating at the pixel level, ViT treats image patches as distinct units of information during training, utilizing self-attention modules to discern their interrelations. ViT has demonstrated superior image classification accuracy over CNNs when ample training data and computational resources are available (Beyer et al., 2022). Nevertheless, the application of vision transformer models for tasks such as weed and crop classification using high-resolution UAV images remains largely unexplored.

In our study, we introduce an innovative methodology for automatically identifying weeds and crops in multispectral images captured by drones, strengthening the vision transformer approach. Our research setup involves a drone equipped with a high-resolution camera, facilitating image acquisition across diverse crop plots under real-world conditions, encompassing tobacco, sunflower, maize, and weed varieties (Czymmek et al., 2019). Our primary aim is to investigate the viability of transformer architectures for specialized tasks like plant recognition in UAV imagery, given the scarcity of labeled data. To address this challenge, we employ data augmentation and transfer learning techniques, supplemented by an evaluation of the self-attention mechanism using vision transformers across varying proportions of training and testing data within a cross-validation framework. Our contributions encompass the integration of low-altitude aerial imagery from UAVs with self-attention algorithms for crop management, pioneering exploration of transformer models for weed and crop image classification, and the assessment of deep learning algorithm generalization capabilities in crop plant classification across different model variations (Alzahrani et al., 2023).

MATERIALS AND METHODS

Study Sites

This study focused on the cultivation of corn, sunflower, and tobacco crops. Drone imagery was obtained from multiple plots in the Kuzmin village of the Sremska Mitrovica region in Serbia during May and June of 2023 (45.0223 N, 19.4052 E) (Figure 1). Table 1 presents the geographical details and agricultural configurations for the crops. An experiment spanning multiple sites was conducted to assess the system's resilience across various ecological zones, aiming to comprehend its adaptability through the diversity of crops and fields involved in the study. The research focuses on industrial crops, with each station subjected to distinct treatment methodologies (Culpan, 2023).

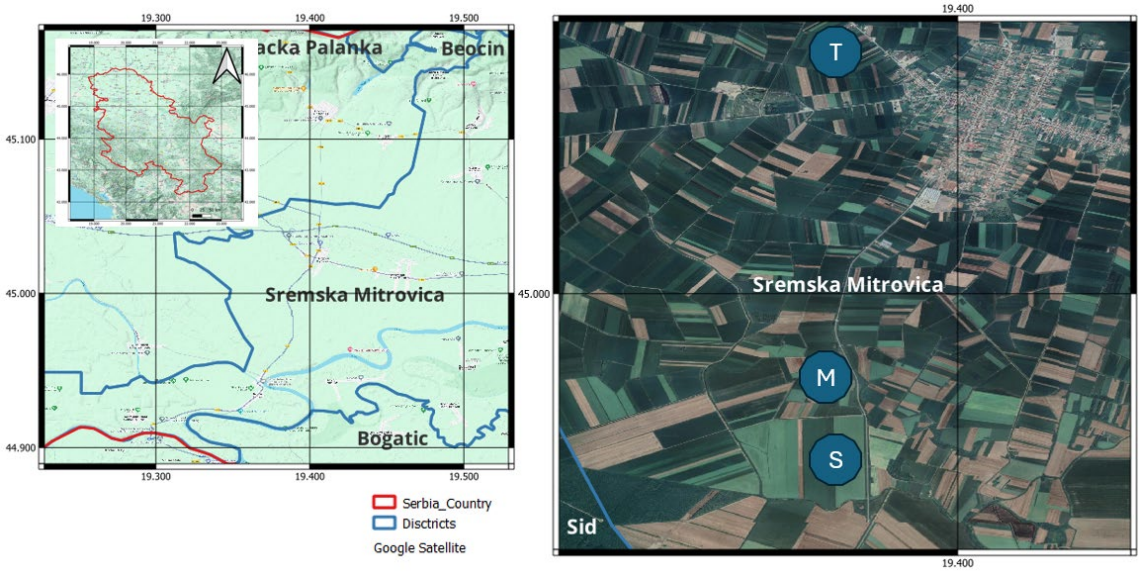

Figure 1. Geographical position of the research area. (S-Sunflower, M-Maize and T-Tobacco, base image from Google Earth)

Table 1. Performance comparison with the state-of-the-models.

	Train Loss	Train Accuracy	Test Loss	Test Accuracy
Swin-B	0.6450	0.9165	0.7256	0.8733
Vit-B16	1.2252	0.9331	1.2516	0.9297
Vit-B32	1.3532	0.9133	1.2878	0.9098

The sunflower images were captured on the 12th (2-leaf stage) and the 14th (4-leaf stage) of June 2023. Photographs of the maize and tobacco were taken during the 3-leaf phase, and on the 8-leaf phase. The density of planting ranges from 33,000 to 45,000 plants per hectare. Soils with excellent filtration capabilities were identified. Additionally, irrigation facilities were available for 95% of the plots, allowing for regulated water conditions (Kayin et al., 2024).

UAV data collection

In this research, imagery was sourced from fields of tobacco, sunflower, and maize, which were sown with inter-row spacings of 60-70 cm. The photographs were taken using a camera mounted on the DJI Mavic 3 multispectral drone. A corpus of 350 RGB photographs was compiled, each with a resolution of 5472 x 3648 pixels and a color depth of 24 bits (Louargant et al., 2017). The environmental conditions during the acquisition of these images were obtained with air temperatures ranging from 24.0 to 26.0 °C and relative humidity levels ranging from 55% to 65%. The drone was equipped with a camera stabilized with a 3-axis brushless gimbal to maintain consistent camera alignment even in strong wind conditions. Flight heights were deliberately set at 10 meters for sunflower plots and 12 meters for maize and tobacco plots. These heights have been optimally chosen to ensure high quality image capture while reducing the duration of drone flights. For maize fields, the increased height was necessary due to more mature plant growth and wing winds. The program of aerial imagery acquisition in different fields was planned to be done at an early stage based on the review and assessment of weed infestation levels in the field. This approach, which spread the imagery over multiple days, resulted in a range of variability and shadow elements in the images, with the tobacco field photographed in the afternoon light conditions, and the sunflower and maize fields photographed in the midday light conditions with the sun at its full perpendicular position.

Prior to launch, the flight path of the UAV was meticulously planned, setting the flight velocity at 2 meters per second. To enhance the quality of image stitching, it was imperative that the overlap of image footprints exceeded 80% both longitudinally and laterally along the flight path. Positional accuracy and altitude control were rigorously maintained within a tolerance of 1 meter and 0.2 meter, respectively, utilizing the Global Positioning System (GPS) and a barometric sensor. Consistent resolution of 0.33 centimeters per pixel was upheld across three different sites, with adjustments in flight elevation compensating for variations in camera pixel sizes to ensure uniform resolution across all imagery.

Image Pre-Processing

For the purposes of this analysis, the model required a comprehensive aerial photograph of the area under investigation. Consequently, prior to conducting any model-based analysis, it was necessary to merge the UAV captured images into a singular, cohesive site map. This image integration process was facilitated using the Pix4Dmapper software (Reedha et al., 2022). All images captured by the UAV were uploaded into the Pix4D software, where the image coordinate system, geolocation data, camera specifications, and other pertinent details were calibrated in alignment with the specifications of the UAV and its camera system (refer to Figure 2).

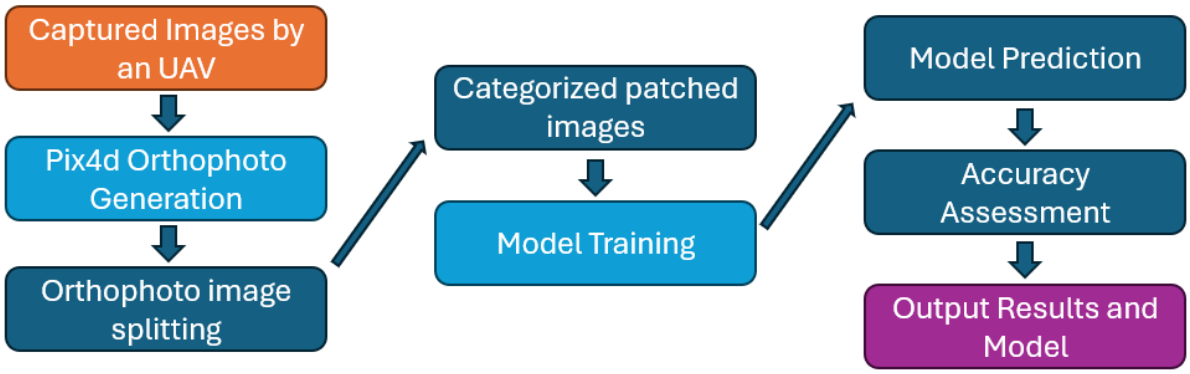

Figure 2. Sequence of steps for image preprocessing and subsequent model forecasting.

From the assembled orthophotos, we extracted specific image segments representing both crops and weeds. These segments were then adjusted to a uniform size of 96x96 pixels. The rationale behind this specific size selection stems from the fact that the dimensions gravitated towards a median of 96 pixels, suggesting a possible proportional relationship between the UAV's operational altitude and the physical scale of the crops documented in the research plots. Weeds were entered into the algorithm by creating classes for weeds, soil and three different agricultural crops.

Vision Transformers (ViT)

Image transducers use the operating principles of transducer models for NLP tasks. It has pioneered a groundbreaking change in deep learning methodology by demonstrating the ingenuity of computer vision. The traditional dominance of convolutional neural networks has been challenged by the success of transformative models in visual data analysis. NLP-based image transformers are attracting the interest of researchers by providing a breakthrough architecture for image classification. The versatility of transformer models also marks a paradigm shift beyond traditional CNN frameworks (Alzahrani et al., 2023).

As described in Vaswani et al's groundbreaking 2017 study, "Attention is All You Need" the core principles of transformers revolve around self-attention mechanisms (Vaswani et al., 2017). As NLP endeavors to demonstrate competence in managing sequential data, the level of importance such mechanisms skillfully assign to different parts of the input data will have increased. These ground-breaking applications of vision transformers apply the self-attention perspective to image inputs and conceptualize them not as conventional pixel grids but as hierarchical arrays of patches, similar to the processing of words in sequential sentences.

Given an image of dimensions H x W x C, the ViT model proceeds by segmenting the images into patches of uniform size, where H represents height, W width and C colour channels, and each patch size is set to have dimensions PxPxC. The total number of patches generated from an image (N) is determined by dividing the total image area by the area of a single patch and is calculated as N = (H/P) x (W/P). For example, for images to be fed at 224x224 with a selected model's patch size of 16, the formula [(224/16) x (224/16)] = [14 x 14] results in a total of 196 patch or array tokens (Xia et al. 2024; Kang et al. 2021).

The way ViT works starts with segmenting images into uniform patches. These patches are then flattened, linearly transformed and enriched with spatial and positional embeddings to preserve spatial and positional context and encoded in a manner similar to the text string processing in transducers. The subsequent stage entails passing the resultant sequence of image patch embeddings through a canonical transform coder architecture (Zhai et al., 2021). The encoder logic consists of multiple layers of multi-head self-attention and feed-forward neural networks, allowing the model to selectively focus on and interpret different regions of the image, thereby distinguishing complex relationships between patches (refer to Figure 3). The output of the transform encoder typically completes the image classification task by generating predictions based on the embedded representations, provided that the output of the transform encoder is fed to the classification head, which typically comprises a linear layer.

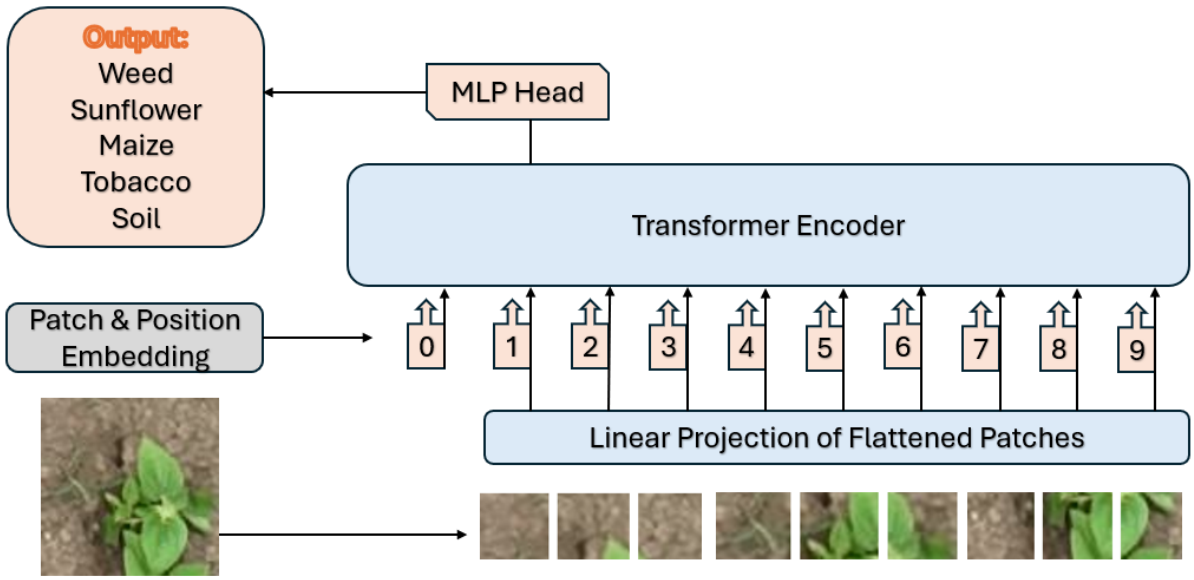

Figure 3. The vision transformer architecture flow shows patch process.

After segmenting the images into patches, the next step is to convert these patches into a one-dimensional format. This transformation can be described mathematically as H*W*C = N*P*P*C. In general, for one-dimensional basis

vector scenarios, the channel value is set to C=1. The flattened patches are mapped to a space equation using a linear transformation layer to produce vectors of size D (Khan et al., 2021). The 'classification token' is created at this stage and passed to the next stage as a class token. The element used as the class token is concatenated with each of the linearly transformed one-dimensional patch vectors and continues, retaining the classification information in the array. The transformer network architecture processes the patches and the class token along with the initial positioning in the array. The learnable embedding created in dimension D is evaluated for use in classification. The methodology of this work reflects the natural language processing strategy of the BERT architecture for image classification within the ViT framework (Bazi et al., 2021).

To preserve the spatial relationships of the original image during the positional encoding process, positional encodings are sequentially added to the patch embeddings. These encodings contain no information about the 2D spatial arrangement of the patches and require the model to learn the spatial relationships between patches from scratch. The transform encoder is fed to the encoder by adding the concatenated patch and position embedding sequence. In the presence of the encoder, the sequence is transformed, allowing the class token introduced at the beginning of the sequence to focus on and assimilate important features from the patches. This process allows a comprehensive embedding process to be learnt, particularly for classification purposes. After the encoder, the class token integrated with the residual information is used to generate a prediction vector by multiplying it with the output of a multilayer perceptron (Han et al. 2020; Suh et al. 2018). After the normalisation layer, the prediction vector becomes capable of image classification using the softmax function and results in a probability distribution. Each image frame is transferred by adding a layer of adaptability and complexity to the architecture of Vision Transformer models at different scales, such as basic and tiny models. The architecture of ViT models, the size of patches, the size multiplicity of embeddings and the self-attention mechanism are often referred to as the width of the model. The depth of the encoder layers, the number of attention heads and the dimensions of the MLP block are called the MLP width and define several basic parameters. These variables allow the ViT model to be customised and optimised for specific task definitions and data sets (Suravarapu et al. 2023; Zhao et al. 2023).

Attention Mechanism

The basic architecture of the converters has enabled significant advances in image analysis through the implementation of query, key and value vectors that are central to their operation. Scaled Dot Product Attention is the core component of this architecture. It allows dynamic weighting of the importance of different parts of the input data. The mathematical formula is defined below as equation (1). Calculates the dot products of queries by scaling them with keys and adding their attention scores. This process improves the model's ability to focus on relevant parts of the data, providing an innovative approach to understanding both textual and visual information (Ma et al. 2023; Mauricio et al. 2023).

$$\text{Self Attention (Q,K,V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)*V \quad (1)$$

In the scaled dot attention mechanism, the variables Q, K and V represent the query vector, transformed key vector and value vector. The scaling process (d_k) is very important for the dimensionality of the key vector. This scaling is done to smooth the dot products and ensure that they remain within the appropriate range. This facilitates a steady gradient decay during the training of deep learning models. The purpose of using (d_k) is to help reduce the potential problem of overly large dot product values that can lead to decreasing derivatives (Shin et al. 2023; Abdalla et al. 2019; Thakur et al. 2023). Preservation of model sensitivity to input subtleties. This balancing act is the basis of attention scoring. This determines how each element in the sequence should manage attention among all the other elements. It increases the model's ability to use relevant information when constructing representations.

In the self-attention process, for each element in the array, the mechanism computes the dot product between the query representation and the key representations of all other entries. The data normalised by applying the softmax operation is transferred to the result set with the attention score, which measures the amount of focus each item should have on all other items in the array.

Since the transformer architecture model assimilates the inputs in a sequential manner, it is through spatial embeddings that parallel data processing is allowed. In encoding the sequence information of the input, embeddings are very important in encoding the sequence information for the transducer to understand and send to the next step (Zhao et al. 2023; Mauricio et al. 2023). The transformer uses its own attention to collectively assimilate the information from each element of the input. It must preserve the order of the data, which is crucial for the operations of the converter, and spatial embeddings must be explicitly added to the input. The positionally enriched inputs are structured into an array with an integrated class embedding based on positional indices to help categorise the input data after the self-attention change (Suravarapu et al. 2023; Ma et al. 2023).

Self-attention works by mapping a set of input vectors onto a set of output vectors, and independently assessing the importance of inputs relative to others. It highlights the importance of context in the process by allowing the model to focus on appropriate aspects of the input. The results of the self-attention module are the sum of aggregated attention scores covering contextual relevance across the sequence. The transformer framework is fundamentally built around these attention mechanisms, often utilizing a multi-head approach to expand the model's capability to focus on various parts of the input simultaneously (Thakur et al., 2023). The scaled dot-product attention and its extension into multi-head attention are pivotal components of this model, which are further elucidated in Figure 4 of the referenced work.

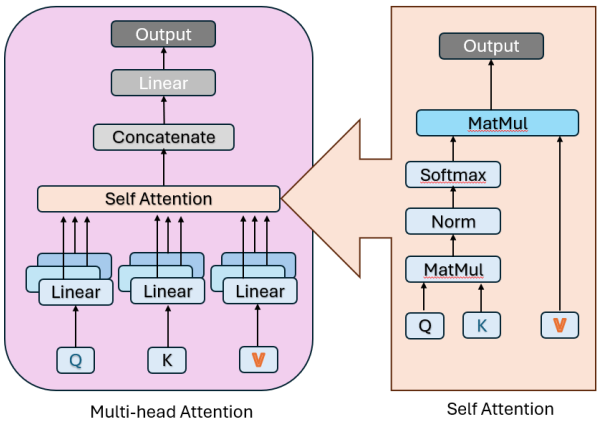

Figure 4. Attention mechanism (Niu et al. 2021).

Evaluation Metrics

We utilized the latest classification methods of Swin B16, ViT-B16 and ViT-B32 models in our project. The research utilized the method of cross-validation to validate the robustness and precision of the proposed models. Cross-validation, which evaluates the performance of a model on test data not used during training, is widely used because of its robustness. It is a method that is attracting attention as a resampling strategy because of its low bias rate. As the classes in our dataset are evenly distributed, we also applied a layered k-fold approach. Ensure that all classes are represented in the validation phase of each fold (Huang et al. 2018; Reedha et al. 2022).

Deep learning models are evaluated by comparing their performance against a benchmark of excellence, known as the gold standard. The accuracy metric is a measure of the model's predictive ability and is calculated as the ratio of correct predictions to the total number of predictions made. (Alzahrani et al. 2023).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

Precision is often used to evaluate the performance of deep learning classification models. Precision indicators are calculated as the ratio of true positive predictions to the sum of true positive and false positive predictions. Precision indicates the model's ability to correctly identify positive specimens among all specimens it classifies as positive (Sunil et al., 2023).

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

$$F1 \text{ score} = 2*\frac{(Precision*Recall)}{(Precision+Recall)} \quad (5)$$

Recall measures the proportion of true positives in the dataset that the model successfully identifies as positive. Simplified, it measures the proportion of true positive predictions made by the model from all positive samples in the dataset. Recall is critical to understanding how effectively the model is able to identify and classify all relevant examples. The F1 score is a widely accepted benchmark for evaluating performance classification scenarios (Hand et al. 2009; Ozcift et al. 2011). It is a method of harmonising and balancing the trade-off between precision and recall by calculating harmonic averages. By integrating both metrics, the F1 score provides a nuanced view of model performance beyond just accuracy.

RESULTS

Each model was trained using a leave-one-out technique with triple cross-validation. The model based on the basic architecture was trained and cross-validated with triple folds and achieved the highest average accuracy of 93.31%. Table 1 illustrates all experimental outcomes related to crop weed classification across the three ViT models. This approach involved training the models on 3150 samples (70%), validating them with 900 samples (20%), and testing them with 450 image samples (10%).

Analysis of the experimental results reveals that the Swin-B16 model surpasses the Vision Transformer models. The Swin-B model achieved the highest accuracy of 91.65% and 0.6450 loss, while the ViT-B16 model closely trailed with a 93.31% accuracy and a minimal loss of 1.2252. All network families exhibit impressive accuracy and F1-score, with the vision transformer models demonstrating the most effective prediction performance in classifying crop and weed images (shown Table 2). The table illustrates the data, revealing notably high recall metric values for each specified category within the dataset. The Swin-B16 model underwent training with batch sizes set at 32, utilizing the SGD optimizer over 20 epochs. Impressively, the model attained an accuracy of 91.65% on both the training and testing datasets. Compared to the other two models, the loss value is notably lower by half, measuring at 0.6450.

The Swin-B16 model exhibited remarkable performance across the first two experiments, displaying consistently high recall and F1-score values. Notably, its performance peaked in the first fold, indicating its efficacy in accurately identifying various classes within the dataset. At the third fold the model showed a decrease in sensitivity, particularly evident in the maize class where sensitivity dropped to 54%. Despite this, the model was successful in discriminating between weed classes and demonstrated the ability to effectively discriminate between different vegetation types.

The ViT-B16 produced consistently impressive results on all three folds. Remarkably, the sensitivity values for the soil class remained high in every experiment, highlighting the robustness of the system in correctly classifying this category. However, the model showed relatively lower sensitivity in the sunflower class compared to other categories. This points to potential difficulties in accurately identifying this crop. It was also observed that the recall for weed classification was relatively low. This indicates some limitations in the accurate detection of weed samples.

Table 2. Evaluated the performance of the models derived from a three-fold cross-validation.

	Experiment 1			Experiment 2			Experiment 3		
Swin-B16	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Maize	0.9700	1.0000	0.9800	0.8900	1.0000	0.9400	0.5400	0.8200	0.6500
Soil	0.9700	1.0000	0.9800	0.9900	0.9400	0.9600	0.6100	0.3000	0.4000
Sunflower	0.9800	0.9200	0.9500	0.9900	0.9600	0.9700	0.7600	0.9900	0.8600
Tobacco	0.9300	0.9700	0.9500	0.9900	0.9800	0.9800	0.7700	0.9400	0.8500
Weed	0.9400	0.8900	0.9100	0.9500	0.9100	0.9300	0.9800	0.4700	0.6300
Vit-B16	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Maize	0.9064	0.9936	0.9480	0.9873	0.9936	0.9904	0.7635	0.9936	0.8635
Soil	1.0000	0.8782	0.9352	1.0000	0.9295	0.9635	0.9923	0.8200	0.9021
Sunflower	0.7919	1.0000	0.8839	0.9286	1.0000	0.9630	0.9398	1.0000	0.9689
Tobacco	0.9792	0.9038	0.9400	1.0000	1.0000	1.0000	1.0000	0.9808	0.9903
Weed	0.9160	0.7692	0.8362	0.9416	0.9295	0.9355	0.9141	0.7500	0.8239
Vit-B32	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Maize	0.8864	1.0000	0.9398	0.8211	1.0000	0.9017	0.9176	1.0000	0.9571
Soil	0.9430	0.6731	0.7749	0.9922	0.8205	0.8982	1.0000	0.8782	0.9352
Sunflower	0.8254	1.0000	0.9043	0.9722	0.8974	0.9333	0.9689	1.0000	0.9842
Tobacco	0.9810	0.9936	0.9873	0.9017	1.0000	0.9483	0.9809	0.9872	0.9840
Weed	0.7465	0.6795	0.7114	0.8889	0.8205	0.8533	0.9032	0.8974	0.9003

As can be seen in the Table 3, the ViT-B32 model showed different sensitivity values in the three folds. On the first fold, the precision values indicate a relatively high classification accuracy, ranging from 74% to 98%. However, in subsequent folds, the precision values varied between 82% and 100% and different values were observed for each class. In particular, the maize and sunflower classes showed excellent recall in all trials. The effectiveness of the model in accurately identifying these specific product types is highlighted.

Table 3. Location and crop planting settings.

Crop Type	Sunflower	Tobacco	Maize
Latitude	19.3873	19.3846	19.3873
Longitude	44.9829	45.0335	44.9829
Date	26/05/2023	22/06/2023	27/05/2023
Flight Height (m)	10	12	12
Row Spacing (m)	0.70	0.65	0.60
Plant Spacing (m)	0.25	0.30	0.25

DISCUSSION

The production of maize and sunflowers is of global importance, with high economic and commercial nutritional value. However, they are susceptible to various diseases and weather events that pose a significant threat to both yield and quality. Early and accurate detection and diagnosis of weed infestations is essential for implementing field-based control strategies and preventing potential losses. This study reaffirms the efficacy of the proposed methods for distinguishing between crops and weeds, offering valuable insights into the performance of various models. The findings of this research demonstrate notable accuracy, laying a solid foundation for the development of automated systems capable of detecting and managing areas affected by weeds in their early stages. Ultimately, such advancements are poised to enhance the efficiency and sustainability of cultivated crop production.

All models underwent training and validation using identical crop samples, encompassing all classes from the same agricultural field. Our investigation reveals that when applied to our agricultural dataset featuring five classes for weed identification, the ViT B-16 architecture, pretrained on the ImageNet dataset, surpasses other architectures and exhibits enhanced resilience to fluctuations in dataset size. Employing ViT for weed classification yields promising outcomes, particularly when dealing with a limited array of classes. In forthcoming experiments, we intend to broaden the dataset by incorporating additional classes to encompass a wider spectrum of crop types. Introducing supplementary classes may potentially lower the classification top-1 score, especially when categorizing plants with analogous shapes and colors. Nevertheless, the ViT is anticipated to yield superior results compared to the Swin-B16 model, given its demonstrated robustness. Although the loss value of the Swin model is lower it cannot provide high precision, recall and F1 scores (Reedha et al. 2022; Wang et al. 2023).

During training, the increments should be applied in such a way as to cover different environmental variations, such as variations in outdoor brightness. The use of augmentations plays an important role in promoting model

convergence and generalisation by changing the examples. This increases the ability of the model to generalise effectively by facilitating the representation of differences in the dataset. If the image acquisition conditions are significantly different, the model's performance may degrade. For example, capturing images of plants after rainfall may result in a change in vibrancy and shape compared to those captured in sunlight. To address these inconsistencies, additional image acquisition is planned for the coming season to ensure robust performance under changing environmental conditions.

CONCLUSIONS

The evolving agricultural environment requires the development of new systems that can accurately identify weeds and crops in different environmental conditions. The solution we used in the classification study overcomes this challenge by exploiting the latest advances in deep learning, pioneered in NLP and now proving useful in computer vision. ViT models have demonstrated superior performance accuracy in a wide range of applications. The model run on the ViT-B16 model using 3 different folding techniques emerges as the best performing model, achieving a test accuracy of 92.97%. Our results also show that the use of smaller patches contributes to improved accuracy. Looking ahead, we aim to develop a hybrid approach to address the complex challenges of crop-weed separation and transformer model design with convolution integration.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: **[https://drive.google.com/file/d/1P8L0V_4-szEByJkODL3TuoNY-H-QxlFx/view?usp=sharing]**.

LITERATURE CITED

Abdalla, A., H. Cen, L. Wan, R. Rashid, H. Weng, W. Zhou and Y. He. 2019. Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure. Comput Electron Agric 167. https://doi.org/10.1016/j.compag.2019.105091

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

Bazi, Y., L. Bashmal, M. M. A. Rahhal, R. A. Dayil and N.A. Ajlan. 2021. Vision transformers for remote sensing image classification. Remote Sensing, 13(3), 516.

Beyer, L., Zhai, X., Kolesnikov, A., 2022. Better plain ViT baselines for ImageNet-1k.

Culpan, E. 2023. Effect of sowing dates on seed yield, yield traits and oil content of safflower in Northwest Turkey. Turkish Journal of Field Crops, 28(1), 87-93.

Czymmek, V., L. O. Harders, F. J. Knoll and S. Hussmann. 2019. Vision-based deep learning approach for real-time detection of weeds in organic farming. In 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) (pp. 1-5). IEEE.

dos Santos Ferreira, A., D.M. Freitas, G.G. da Silva, H. Pistori, M.T. Folhes. 2019. Unsupervised deep learning and semi-automatic data labeling in weed discrimination. Comput Electron Agric 165, 104963. https://doi.org/10.1016/J.COMPAG.2019.104963

dos Santos Ferreira, A., D. Matte Freitas, G. Goncalves da Silva, H. Pistori, M. Theophilo Folhes, 2017. Weed detection in soybean crops using ConvNets. Comput Electron Agric 143, 314–324. https://doi.org/10.1016/j.compag.2017.10.027

Han, K., Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang and D. Tao. 2020. A Survey on Visual Transformer. https://doi.org/10.1109/TPAMI.2022.3152247

Hand, D. J. 2009. “Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve.” Machine Learning 77 (1): 103–23. doi:10.1007/s10994-009-5119-5.

Hasan, A.S., M.M., F. Sohel, D. Diepeveen, H. Laga, M.G.K. Jones. 2021. A survey of deep learning techniques for weed detection from images. Comput Electron Agric. https://doi.org/10.1016/j.compag.2021.106067

He, K., X. Zhang, S. Ren and J. Sun. 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

Huang, H., J. Deng, Y. Lan, A. Yang, X. Deng, S. Wen, H. Zhang and Zhang, Y., 2018. Accurate weed mapping and prescription map generation based on fully convolutional networks using UAV imagery. Sensors (Switzerland) 18. https://doi.org/10.3390/s18103299

Iqbal, N., S. Manalil, B.S. Chauhan and S.W. Adkins. 2019. “Investigation of Alternate Herbicides for Effective Weed Management in Glyphosate-Tolerant Cotton.” Archives of Agronomy and Soil Science 65 (13). Taylor and Francis Ltd.: 1885–99. doi:10.1080/03650340.2019.1579904.

Iqbal, N., S. Manalil, B.S. Chauhan, S.W. Adkins. 2019. Investigation of alternate herbicides for effective weed management in glyphosate-tolerant cotton. Arch Agron Soil Sci 65, 1885–1899. https://doi.org/10.1080/03650340.2019.1579904

Kang, J., L. Liu, F. Zhang, C. Shen, N. Wang, L. Shao. 2021. Semantic segmentation model of cotton roots in-situ image based on attention mechanism. Comput Electron Agric 189. https://doi.org/10.1016/j.compag.2021.106370

Kayin, G.B., H. Kayin, A.T. Goksoy. 2024. Effects of Plant Density on Micronutrient Uptake in Sunflower (*Helianthus annuus* L.) Varieties. Turkish Journal of Field Crops 29, 9–17. https://doi.org/10.17557/tjfc.1349344

Lecun, Y., Y. Bengio, G. Hinton. 2015. Deep learning. Nature. https://doi.org/10.1038/nature14539

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

Louargant, M., S. Villette, G. Jones, N. Vigneau, J.N. Paoli and C. Gée. 2017. Weed detection by UAV: simulation of the impact of spectral mixing in multispectral images. Precis Agric 18, 932–951. https://doi.org/10.1007/s11119-017-9528-3

Lu, Y. and S. Young. 2020. A survey of public datasets for computer vision tasks in precision agriculture. Comput Electron Agric. https://doi.org/10.1016/j.compag.2020.105760

Ma, H., L. Zhao, B. Li, R. Niu, and Y. Wang. 2023. Change Detection Needs Neighborhood Interaction in Transformer. Remote Sens (Basel) 15. https://doi.org/10.3390/rs15235459

Madsen, S.L., S.K. Mathiassen, M. Dyrmann, M.S. Laursen, L.C. Paz and R.N. Jørgensen. 2020. Open plant phenotype database of common weeds in Denmark. Remote Sens (Basel) 12. https://doi.org/10.3390/RS12081246

Maurício, J., I. Domingues and J. Bernardino. 2023. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Applied Sciences (Switzerland). https://doi.org/10.3390/app13095521

Niu, Z., G. Zhong and H. Yu. 2021. A review on the attention mechanism of deep learning. Neurocomputing 452, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

Ozcift, A. and A. Gulten. 2011. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput Methods Programs Biomed 104, 443–451. https://doi.org/10.1016/J.CMPB.2011.03.018

Radoglou-Grammatikis, P., P. Sarigiannidis, T. Lagkas and I. Moscholios. 2020. A compilation of UAV applications for precision agriculture. Computer Networks 172, 107148. https://doi.org/10.1016/J.COMNET.2020.107148

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

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

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

Sunil, C.K., C.D. Jaidhar and N. Patil. 2023. Systematic study on deep learning-based plant disease detection or classification. Artif Intell Rev 56, 14955–15052. https://doi.org/10.1007/s10462-023-10517-0

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

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

Ustuner, T., A. Sakran and K. Almhemed. 2020. Effect of Herbicides on Living Organisms in The Ecosystem and Available Alternative Control Methods. International Journal of Scientific and Research Publications (IJSRP) 10, 622–632. https://doi.org/10.29322/ijsrp.10.08.2020.p10480

Thakur, P.S., S. Chaturvedi, P. Khanna, T. Sheorey and A. Ojha. 2023. Vision transformer meets convolutional neural network for plant disease classification. Ecol Inform 77. https://doi.org/10.1016/j.ecoinf.2023.102245

Vaswani, A., G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser and .I. Polosukhin, 2017. Attention Is All You Need.

Vrbničanin, S., D. Pavlović and D. Božić. 2017. Weed Resistance to Herbicides, in: Herbicide Resistance in Weeds and Crops. InTech. https://doi.org/10.5772/67979

Wang, H., W. Chang, Y. Yao, Z. Yao, Y. Zhao, S. Li, Z. Liu and X. Zhang. 2023. Cropformer: A new generalized deep learning classification approach for multi-scenario crop classification. Front Plant Sci 14. https://doi.org/10.3389/fpls.2023.1130659

Wu, X., S. Aravecchia, P. Lottes, C. Stachniss and C. Pradalier. 2020. Robotic weed control using automated weed and crop classification. J Field Robot 37, 322–340. https://doi.org/10.1002/rob.21938

Xia, Z., X. Pan, S. Song, L. Erran Li and G. Huang, 2022. Vision Transformer with Deformable Attention.

Zhai, X., A. Kolesnikov, N. Houlsby and L. Beyer. 2021. Scaling Vision Transformers.

Zhao, J., T.W. Berge and J. Geipel. 2023. Transformer in UAV Image-Based Weed Mapping. Remote Sens (Basel) 15. https://doi.org/10.3390/rs15215165