AI-Enabled Vision Transformer for Automated Weed Detection: Advancing Innovation in Agriculture

Shafqaat Ahmad, Zhaojie Chen, Aqsa, Sunaia Ikram, Amna Ikram
Data Scientist, Brandt Group of Companies, Canada
Department of Food Science and Nutrition, The Hong Kong Polytechnic University
Department of Computer Science, COMSATS University Islamabad, Pakistan
Department of Software Engineering, IUB, Pakistan
Department of Computer Science, GSCWU, Pakistan

***Abstract*—Precision agriculture is focusing on automated weed detection in order to improve the use of inputs and minimize the application of herbicides. The presented paper outlines a Vision Transformer (ViT) model for weed detection in crop fields, that tackle difficulties stemming from the resemblance of crops and weeds, especially in complex, diversified settings. The model was trained via pixel-level annotation of the images obtained using high-resolution UAV imagery shot over an organic carrot field with crop, weed, and background. Due to the nature of the mechanism in ViTs that includes self-attention, which allows it to capture long-range spatial dependencies, this approach can very well distinguish crop rows from inter-row weed clusters. To solve the problem of class imbalance and improve the generality of the patches, techniques of data preprocessing such as patch extraction and augmentation were used. The effectiveness of the proposed approach has been confirmed by an accuracy of 89.4% in classification, exceeding the efficiency of basic models such as U-Net and FCN in practical application conditions. This proposed ViT-based approach is a marked improvement in crop management; and provides the prospect for selective weed control, in support of more sustainable agriculture. This model can also be integrated into AI-based tractors for real-time weed management in the field.**

Keywords*—*Precision agriculture; weed detection; vision transformer; UAV imagery; crop-weed classification; AI-Tractors

I. Introduction

In light of such global factors as climate change, increasing population, and declining land fertility, protection of food production has become an important task [1]. Amongst various biotic constraints that affect crop yield and quality, weeds rank as some of the most formidable challenges that crop producers face in the field [2]. If not controlled, weeds have severe effects on crop yield and quality hence contributing to loss making and high food insecurity [3]. In the past, weed management has been undertaken by mechanical means such as pulling weeds out by hand or the widespread application of herbicides [4]; either of which is now considered to be unfavorable. Hand weeding is hard and cannot be used in large scale farming [5], while chemical control causes pollution and health issues [6], reduces bio-diversity, and results into the evolution of herbicide resistant weeds. Therefore, there is need to develop effective, sustainable as well as economic methods to tackling weed problems.

New developments in precision agriculture especially in combination with technology such as remote sensing, machine learning and drone systems, are revolutionizing conventional weed control approaches [7]. Precision agriculture is the practice of trying to grow crops as efficiently as possible by giving farmers instant information about the condition of their fields so they can manage the resources they use in the most sustainable way [8]. With UAVs using high resolution and multispectral cameras available for field monitoring, large scale data acquisition coupled with detailed visualisation of crop and weed distribution in the agricultural environments is possible [9]. It is possible to use this technology to locate and identify weeds and subsequently manage by providing efficient spot treatment as opposed to weed eradication using herbicides.

Nevertheless, identification and precise categorization of weeds in crop fields still pose a great challenge due to factors such as variability in the field, weeds growing between rows of crops and close resemblance in appearance of weeds and crops [10]. These difficulties cannot be resolved by using conventional image processing techniques, because such approaches rely on color-based or shape-based segmentation, which may not be sufficient for distinguishing between very similar plant species in different lighting and environmental conditions [11]. To overcome these limitations, machine learning particularly deep learning approaches has been used to improve weed detection accuracy. Convolutional neural networks (CNNs) have been reported to work well in this area [12], however, due to their constrained local connectivity, they lack the ability to capture the spatial dependencies and context required to correctly identify weeds from crops especially in high density field setting.

Recently, Vision Transformers (ViTs) emerged as a compelling approach to surpass CNNs in image classification problems [13]. First introduced for natural image understanding, ViTs utilize the self-attention method and it provides a wide-angle view of long-range dependencies within the image, which is crucial in agriculture. Unlike CNNs, ViTs can handle the analysis of the entire image regions rather than focusing on localized features needed for crop and weed differentiation [14]. Self-attention enables ViTs to distinguish between crop rows and inter-row weed clusters more accurately than in field conditions where crop plants and weeds appear to have similar textures and color patterns.

This research introduces a new method for the automated detection of weeds based on a Vision Transformer model that has been developed to handle the specific difficulties of agricultural weed categorisation in UAV imagery. The proposed method takes advantage of the fact that crops, as a rule, are planted in a geometric pattern of rows while weeds grow randomly across the farm field; therefore, crop regions can be distinguished from the clusters of weeds by their geometric arrangement. The method we propose here is to employ the Vision Transformer model on the high-resolution UAV images at the pixel level to accurately distinguish crops from weeds. The training dataset is CWFID, for each image, background, crop, and weed pixels are labeled in detail with the help of experienced farmers, which supplies the model with profound features for learning intricate spatial associations.

Using the efficiency of image preprocessing including patch extraction and data augmentation and the feature of long-range dependencies analysis of ViT model we expect to receive high classification accuracy and good scalability in field conditions. This study advances understanding of weed biology and the potential for selective, efficient weed control by identifying specific proteins that allow for accurate discrimination of different weed species. Consequently, the study responds to important research questions in PA and opens up opportunities toward building more sustainable and less hazardous crop growing systems.

The remainder of this paper is organized as follows: Section II discusses related work, which presents an idea of this research area and the inclusive techniques for weed detection and its merit and demerit. Section III outlines the research approach of this study, which covers ViT architecture, datasets, data preprocessing, and evaluation of crop and weed classification. Section IV explains the findings that include the assessment of the ViT model and a comparison with other conventional models like U-Net and FCNs. Section V offers a discussion of the findings, issues on model stability and possible applications of the developed models to precision agriculture. Last, Section VI provides a conclusion to the study by offering an overview of the major conclusions, the main research contributions, and an indication of the areas where future studies and enhancements may be made.

II. Related Work

There are different approaches for weed detection mentioned in the literature for the use of different image acquisition systems. The first one is carried out by separating vegetation from the background as soil and residues to separate crops from the weeds. The common segmentation process handily uses the color Methods [15] and Multispectral data in order to segment vegetation from background using fixed indices which make vegetation segmentation possible. Nonetheless, differentiating between weeds and crops using spectral data prove difficult since the two are spectrally similar. Therefore, approaches focusing on the region level, which utilizes spatial pixel configurations, are mostly used [16].

The detection of weeds in agriculture has improved over the years with the help of color-based segmentation algorithm. Hue based indices like the Excess green Index (ExG) are used widely to sharpen vegetation features in imagery, isolating plants from their surroundings. This approach is especially valuable when dealing with multispectral data since, as it was mentioned, ExG uses the green component most to enhance vegetation. This method has been found to be computationally efficient for the initial step of separating crops from weeds in agricultural scenarios and laid down a base for further analysis and classifying more steps [17]. Another level of enhancement of weed detection is obtained by integrating Excess Green with Otsu's thresholding technique which segment images at the optimum threshold intensity values. The integration method is passes in minimizing the background noise while maximizing vegetation details. Together with the double Hough transform, this method improves the identification of crop lines in images with perspective distortion by recognizing and reorienting the lines in a complex environment in agriculture. They are particularly useful in the images of the same scene taken under varying lighting conditions since they increase the resistance when classifying crops from weeds [18].

TABLE I. Previous Weed Detection Methods

Method	Description	Reference
Color-Based Segmentation	Separates vegetation from background using color indices such as Excess Green (ExG) and fixed indices in multispectral data.	[17]
ExG and Otsu's Thresholding	Combines Excess Green and Otsu's thresholding to eliminate background, then uses double Hough transform to identify crop lines in perspective images.	[18]
Object-Based Image Analysis (OBIA)	Uses UAV imagery and multiscale algorithms to segment crop rows from weeds, creating homogeneous objects for analysis.	[19]
2D Gabor Filters with ANN	Uses 2D Gabor filters to capture texture features and an artificial neural network (ANN) classifier for weed detection.	[20]
Morphological Characteristics	Utilizes morphological features to distinguish weeds in maize fields, using neural networks and support vector machines (SVMs) with shape-based features.	[21]
Edge Frequencies & Vein Density	Differentiates weeds from crops by analyzing edge frequencies and vein density differences in the leaves.	[22]
Otsu Thresholding & K-means/SVM	Applies Otsu thresholding for background removal and uses k-means clustering and SVM classifier for crop-weed classification, successful in sunflower fields.	[12]
Wavelet Transform & Neural Network	Uses wavelets to capture texture details and a neural network for classification, effective for recognizing various weed types in sugar beet fields.	[23]
SVM, ANN, & Random Forests in OBIA	Employs machine learning models like SVMs, ANNs, and Random Forests within the OBIA framework, especially for weeds in maize fields.	[24]
Convolutional Neural Networks (CNNs)	Uses CNN architectures, including AlexNet, for weed detection in crops such as water hyacinth and serrated tussock. Applied in UAV-based imagery and mobile robot systems.	[25]
Spatial & Spectral Domain Features	Integrates Hough transform for spatial features with multispectral data for spectral features, combined with SVM for crop-weed classification in four-band imagery.	[26]

Another complex technique is Object-Based Image Analysis (OBIA), which divides images into areas of the same character instead of single pixels, within the use of multistate algorithms. When applied to UAV imagery, OBIA provides a better defined and can be easily automated procedure to distinguish crops from weeds. This approach is useful in vast areas where exact methods like pixel based approach turn out to be more computational. Thanks to OBIA, grouping similar pixels into coherent objects, researchers are able to distinguish the pattern of weed distribution across the crop rows, which enhances weed control strategies [19]. The combination of texture analysis in the form of 2D Gabor filters with Artificial Neural Networks (ANNs) introduces a promising solution to the problem of weed detection due to the utilization of frequency and orientation within images. The enhanced textural features that are fundamental to crops and weeds are well captured by Gabor filters. ANNs then sort these features, and the model has a high level of accurate weed detection in crops with textural features. This method offers considerable reliability to precision agriculture, above all in areas of uniform textural characteristics where texture differential is significant [20].

Shape based features of Morphological characteristics are another factor that builds another level of discrimination in case of weed identification. Methods that apply such factors as shape, size, and structure of the leaves using neural networks and Support Vector Machines (SVMs) are preferable in structured crops such as maize. Morphological features are unique depending on the type of crop or the weed in question, and therefore helpful where the shape differences are quite profound. Such a strategy can be especially valuable for detecting specific weed types that differ from crops morphologically [21]. Another notable feature which is used in classification of weeds is the patterns that appear on the leaves 'veins. Vein density methods and edge frequency methods help to distinguish crops and weeds because crops and weeds essentially have different vascular networks within the veins of their leaves naturally. This technique is most successful in the controlled environments where crops and the weeds differences in vein densities are clearly noticeable. Due to this focus on these several anatomical dissimilarities, this approach is suitable for high precision detections in small-scale or research production agriculture setting [22].

Furthermore, using thresholding Otsu together with clustering and classification method such as K-means and SVM makes a strong way of detecting weeds in areas such as sunflower crops. Otsu's thresholding erases the background noises while k-means clusters all the pixels having a nearly similar intensity, which is then sophisticatedly classified by the SVM in order to separate weeds correctly. Thus, this work follows gradual layering of steps that help increase the weed detection accuracy, and that are tested effective even in high noise images [12]. When used alongside neural networks, the wavelet transform is a useful method of weed detection through texture analysis. Wavelets analyze small local details of the image and since neural networks can provide high accuracy when determining the difference between the weeds. This technique has been particularly effective in sugar beet fields where due to the multi specie flora the different weeds can be identified using the features obtained by the wavelet analysis of the images [23].

Currently, the use of OBIA has included some common machine learning models, such as SVMs, ANNs, and Random Forests. This approach especially for maize fields incorporates an object-based image analysis with machine learning concept leading to higher accurate detection in large-scale agriculture. Thus, the classifiers within and across the imagery segments enhance the models to increase classification results in high complexity areas where the mere pixel-based approach could not limit the classification process [24]. Convolutional Neural Networks (CNNs) are that key technology which helps weed detection using high-dimensional data and pattern extraction. The state of the art CNNs, such as the AlexNet, has been implemented in the classification of weed crops such as water hyacinth and serrated tussock. These models are particularly suitable for UAV and mobile robotic systems where high versatility of weeds and constant ability to perform well in different conditions is needed. The feature extraction capacity of CNNs makes them useful in agricultural systems particularly where big data samples can be used in training and model refinement [25].

Last of all, advanced techniques that combine spatial and spectral characteristics of the analysed images, including Hough transform method with the use of multispectral imaging and support vector machines, can be pointed to as an enhanced method for crops and weeds differentiation. This approach takes advantage of spatial characteristics and spectral variation of four bands in imagery for precise analysis in precision agriculture. This method involves combining of spectral data with spatial transformation to result in high classification accuracy particularly in fields where spectral and spatial discrimination is well defined [26].

III. Proposed Methodology

In modern agriculture, most crops are planted in organized rows with defined spaces, depending on the crop type. Vegetation that grows outside these rows is generally identified as weeds, known as inter-row weeds. Leveraging this spatial organization, several studies have implemented weed detection methods based on the geometric properties of crop rows. A key benefit of these methods is that they are largely unsupervised, reducing the need for manual training data. Building on this, our approach first identifies crop rows, then labels inter-row vegetation as weeds to create a training database. We categorized this data into two classes, crop and weed, and used it to train a Vision Transformer (ViT) model to detect and classify crops and weeds from UAV imagery. Fig. 1 provides an overview of the main steps in the proposed method, with detailed descriptions following.

Crop/Weed Field Image Dataset (CWFID) is one of the vital resources for demonstrating the models of machine learning for classification of crops from weeds. The data in this paper was obtained from an organic carrot field in Northern Germany with the help of an autonomous field robot called Bonirob which has a high-resolution multi-spectral camera. Collected during the vegetation phase of the crops, the images offer a real-world representation of crop and weed status in the fields, which is useful for precision agriculture studies with detailed descriptions of both crops and weeds present in the image. The dataset comprises 60 high-resolution images with the size of 1296 x 966 pixels. The fine details present in the presented images make it possible for models to differentiate vegetation features, and also differentiate between plants that are growing closely together. Each image in the dataset is fully annotated at the pixel level by agricultural experts, classifying each pixel into one of three categories: The three categories of organisms identified in the study area include Background (Soil), Crops (Carrot Plants) and Weeds.

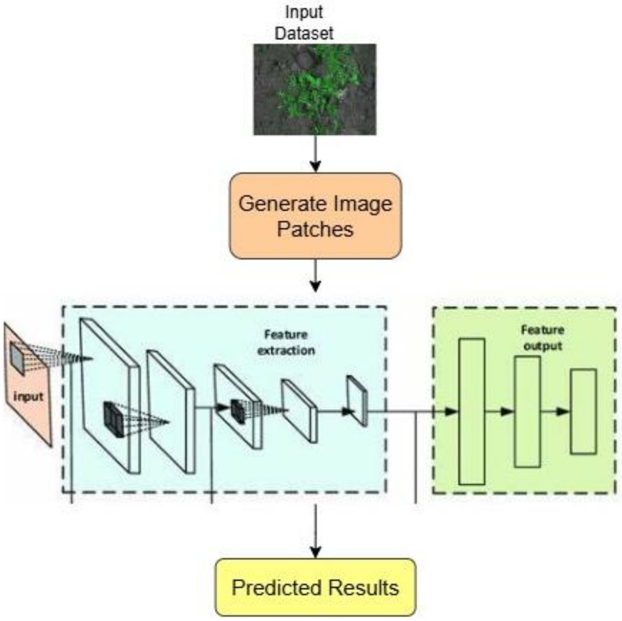

Fig. 1. Proposed architecture flow.

The expert annotations of the CWFID dataset allow for the identification of crops and weeds in a highly accurate, even in the most complicated agricultural environment. Every class in the case of the given dataset is a category of content that helps in differentiating between plants and soil. The three annotated classes are described in detail below:

1) Background (Soil): This class consists all the bare grounds particularly the soil that is inter and intra-row of crops. These background regions makes it easier to define whether an object is crop or weed since they creates a contrast. The pixel distribution across the three classes is as follows and has been presented in Table II to indicate the class imbalance problem similar to that of realistic problem.

2) Crops (Carrot Plants): This class is made up of areas with carrot plant bodies, which are usually aligned in an orderly manner. They contribute to the improvement of the effectiveness of the classification between crop and non-crop areas due to the crop rows formed. The carrot plants were in different stages of maturity which provided a variety that helps the models understand different crop morphologies.

3) Weeds: In this dataset, weeds comprise intra-row weeds, which are those established within the crop rows, and inter-row weeds, which are those established between the crop rows. Classification models are further complicated by the presence of intra-row weeds since they are similar in size and color to crops. The presence of a large number of different weeds improves the applicability of the dataset for developing reliable machine learning models capable of distinguishing between crops and different weed types.

Because of the high-resolution images and corresponding detailed annotation, the CWFID dataset is more suitable for precision agriculture in which precise weed maps are required at the pixel level. The dataset offers several unique challenges:

1) Class imbalance: The fact that there are more background and crop pixels than weed pixels is a more realistic representation of the field environment to encourage the use of methods such as data enhancement and class balancing.

2) Intra-row and inter-row weeds: Moreover, the combination of both inter-row and intra-row weeds poses a difficult classification problem for the models, where the presence of weeds in between crop rows is also considered.

3) Varied lighting and vegetation density: The dataset comprises images taken under varying light conditions and at different vegetation cover densities making it more challenging to classify while improving model resilience.

The CWFID dataset is freely accessible from GitHub and can be used by researchers interested in crop and weed classification for agricultural applications. The specific use of this dataset is to contribute to the generation of models for improving precision agriculture, especially in the case of weed detection and control within crop fields.

A. Data Preprocessing

To train effectively and to generalise the crop-weed classification model, some modifications were made on the CWFID dataset during data preprocessing. These steps were taken in an effort to bring the format of the data fed into the classifier more to a unified level, also to equalize the ratio of the classes and increase the variety of training samples (Table I).

1) Image resizing: When analyzing the CWFID dataset, it was found that each of the original images has a size of 1,296 x 966 pixels, and thus requires downscaling for input into the most of the deep learning models. To keep it manageable for the model, all of the images were also scaled to a size that would fit the input size of the chosen model. This resizing made all the inputs have equal dimensions thus making the model to undergo training without the need for further resizing during training. The option of resizing was applied more conservatively, allowing the image to maintain as much of its quality as possible, and at the same time, decreasing the amount of computations needed.

2) Patch extraction: To prepare the data for pixel level classification, each resized image was then split into fixed size patches of 8×8 pixels. Patch extraction serves several purposes:

a) Localized feature capture: Since a large image is divided into small segments of patches, C&Ps can identify crop and weed characteristics confined or restricted to particular regions.

b) Class representation: Each patch was made around regions marked as crop or weed and both of them were equally represented in the training set.

c) Memory efficiency: Smaller patches also mean that Memory is used, which is a great advantage since more patches mean more data for the AI model to train on, especially and particularly when using humongous data sets.

In general, patches consisting only of background pixels were often removed in order to focus on crop and weed area. They also diminished unnecessary data and at the same time enhanced the specificity of the data set with regard to crops and weeds differentiation. If a patch contained both crop and weed pixels it was classified into a patch class with the highest pixel count in a particular patch in order to of class labeling.

3) Data augmentation: Since there are significantly fewer weed pixels than crop pixels in the CWFID dataset, data augmentation was used to increase the number of weed samples and, therefore, improve the model's performance. Augmentation techniques were applied uniformly across both classes to expose the model to a variety of conditions, as detailed below:

a) Rotation: Each patch was rotated at 90°, 180°, and 270° rotations. This rotation not only amplified the dataset up to four folds but also let the model learn about the rotational variability needed for recognizing weeds and crops in multiple angles.

b) Contrast adjustments: To mimic different lighting scenarios that could be met in practice, the contrast of a patches was altered randomly. Crop, weed, and background boundaries were highlighted through higher contrast settings; lower contrast mirroring conditions such as low light or shadow. This adjustment improved the ability of the model to be sensitive to variations in the levels of illumination in the environment.

c) Gaussian smoothing: Specifically, Gaussian smoothing, or blurring, was applied only to minimize the noise in the image and enhance the main characteristics of each patch. High frequency components and significant intensity variations were removed through applying a Gaussian filter, and this enabled the model to detect general features. This technique also assisted in lowering the impact of noise and enhanced generalization in some instances.

4) Balanced dataset composition: To reduce the class imbalance, dataset was augmented in such a way that both crops and weeds had almost equal representation. Augmented weed patches were particularly helpful in countering this in data collection because crop areas are generally more abundant. To this end, the findings demonstrated that it is possible to get a near balanced distribution between the two classes and this made the model to perform well in making discriminations between the two classes without favoring the larger class. Through this detailed data preprocessing step, the CWFID dataset was well-prepared for training the Vision Transformer model, which then captured important aspects of both crops and weeds and succeed under different field conditions.

TABLE II. SUMMARY OF DATA PREPROCESSING TECHNIQUES USED

Preprocessing Step	Description	Purpose
Image Resizing	Standardized input size for all images	Ensures consistency and reduces memory usage
Patch Extraction	64 × 64 pixel patches centered on crop or weed regions	Localizes features and increases efficiency
Rotation	Rotations at 90°, 180°, and 270° angles	Increases data size and rotation invariance
Contrast Adjustments	Simulates lighting variations by adjusting contrast	Improves robustness to different lighting
Gaussian Smoothing	Applies a Gaussian blur to reduce noise and enhance primary features	Focuses model on main features, reduces noise

B. Model Training

The prepared CWFID dataset was used to train a Vision Transformer model because of its efficiencies in capturing the spatial relationships within the image data. In contrast to the standard convolutional models, the ViT model adapts a self-attention mechanism, enabling the model to acquire contextual data from larger regions of each picture, which makes it suitable for learning subtle distinctions between crops and weeds. To evaluate the model's performance effectively, the dataset was split into separate training and testing sets. Eighty percent (80%) of the images were allocated to the training set, with the remaining 20% reserved for testing. This split ratio was chosen to ensure that the model could learn robustly from a substantial amount of data while still providing a sufficient amount of unseen data for accurate performance evaluation.

Care was taken to maintain a balanced distribution of crop and weed samples within both sets, allowing the model to be tested on images that represent the diversity and complexity of real-world conditions captured within the CWFID dataset. This split provided the model with an appropriate balance between learning general features during training and evaluating its effectiveness in generalization during testing. To optimize the ViT model for the crop-weed classification task, a set of training parameters was carefully selected based on preliminary testing and validation:

1) Optimizer and learning rate: The function used for optimization was presented by the stochastic gradient descent (SGD) with the learning rate equal to 0.001. It was chosen due to its performances in dealing with large number of sample inputs and the fact that it can converge significantly when trained with appropriate learning rate. The learning rate of 0.001 was found to give a stable and systematic training improvement to the model without oscillating training or causing a convergence problem.

Algorithm: Vision Transformer (ViT) for Crop and Weed Classification

```
Input:
    • Image dataset D with labeled crop, weed, and background images
    • Pre-trained Vision Transformer model ViT
    • Training parameters: batch size, learning rate, number of epochs
Output:
    • Classified images with crops and weeds distinguished
Initialization
    1. Load images from dataset D and associated labels (crop, weed, background). 1.2 Apply data transformations to each image:
        – Resize to 224×224= times (ViT input size).
        – Apply random horizontal flip, rotation, and normalization.
    2. Define a custom dataset class CropWeedDataset for loading images and labels.
    3. Initialize DataLoader for training and validation datasets with the transformed images.
    4. Initialize the Vision Transformer model ViT with a classification head suitable for the number of classes
    5. Set the loss function as Cross-Entropy Loss
    6. For each epoch in the specified number of epochs: - Set the model to training mode.
    7. Perform backpropagation and update model weights
    8. Perform a forward pass through the model. - Compare predictions to actual labels to calculate accuracy.
```

2) Batch size: The batch size of 16 was chosen, as such size is more efficient in terms of memory and computation speed. This size ensured the model could handle a reasonable amount of data per every step, the training and convergence process was much smoother and quicker compared to the larger batch sizes, but the memory issues that can come with large batch sizes were also avoided.

3) Epochs: In initial experiments, it was defined that the number of epochs should be 50. This decision was made based on observing the loss and accuracy plots during trial runs of the model several times and noted that 50 epoch was sufficient for the model to learn the features required to distinguish crops from weeds without over-fitting to the training data. Of note, early stopping and validation checks were used to stop training if the model was overfitted or if the training process stagnated, for purposes of time and computational efficiency.

a) Training process: During training, the ViT model took in each 64 × 64 pixel patch derived from the CWFID dataset. The self-attention within the ViT structure allowed the model to learn spatial relationships among these patches thus distinguishing between crop and weed patterns well. Maintaining constant observance of the training and validation loss made it possible to check if the model is overfitting or underfitting. When this training setup was complemented with the well-prepared dataset and the augmentation strategies, the ViT model was able to generalize well. Upon the completion of training, the model was able to learn different patterns and spatial relationship of crops and weeds for a robust classification during the test. Fig. 2 shows ViT architecture.

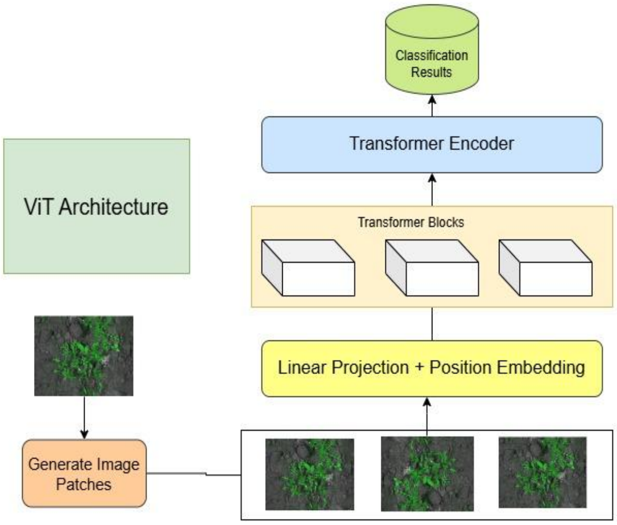

Fig. 2. ViT Architecture.

IV. Experimental Setup

For this study, a systematic experimental framework was developed to assess the ViT approach for crops–weeds discrimination based on the CWFID dataset. This setup entailed setting not only the hardware but also the software environment to address the requirements of processing high-resolution UAV imagery and running deep learning models such that we would obtain reproduceable results.

The experiments were performed on a high-performance computing system consisting of an Intel Xeon E5-2678 v3 processor (2.5GHz), and an NVIDIA GeForce GTX 1080 Ti GPU with 11 GB VRAM. This combination of CPU and GPU allow to process large image datasets effectively and speedup model training. In the framework of the proposed system, it utilized 64GB of DDR4 RAM which make use of the in-memory data processing, especially helpful when dealing with a significant volume of augmented samples. A 1TB SSD was used to store the dataset and the intermediate outputs so that during training and evaluation phase, there was low latency and fast data access.

Regarding the software environment the experiments were performed on Ubuntu 20.04 LTS operating system because of its ability to support deep learning frameworks and successfully manage computationally intensive tasks. PyTorch 1.9.0 has been the major library used to train the ViT model since it offers flexibility to implement transformer models. Furthermore, basic Python libraries including OpenCV for image processing, NumPy for numerical computation and scikit-learn for assessing the performance of the offered models were also installed into the environment. The transformation of the images was made possible by using the Albumentations library towards the enhancement of the data augmentation processes so as to enhance sample variability.

The dataset preparation for the current study followed the preprocessing steps as outlined in the methodology. The CWFID dataset was further split into a training set and a testing set out of which 80% was used for training and the rest was used for testing The training and testing datasets contained equal numbers of crop and weed samples. This division allowed for providing the model with enough samples for learning the general features at the same time preserving a separate part for the model accuracy assessment of the new, previous samples. Every single high-resolution image was partitioned into 64 × 64 pixel window patches that were centered on crop or weed annotations. This patch extraction enabled the decoupling of complex features with this model to learn localized features while data augmentation including rotation, contrast adjustment and Gaussian smoothing were used to increase the variation in lighting, orientation and appearance of crops and weeds samples.

Some of the parameters of the ViT model were set specifically for the purpose of crop and weed classification. Its architecture was based on the self-attention mechanism and was selected due to its capability to define spatial dependencies within the patches of images successfully. The model was trained with following specifications: batch size = 16, learning rate = 0.001 and the optimizer used for training is stochastic gradient descent. The total training process comprised 50 epochs, if validation loss stopped increasing or began to rise, early stopping was used to stop training. The cross entropy loss was adopted as the main loss function, which provides flexibility in multi-class crop, weed, and background classification.

To assess the performance of the developed models, a variety of evaluation measures was used. The accuracy for each of the classes, namely the crop, weed and background were determined in order to compare the performance of the models. With accuracy and quantity, measures of precision and recall were useful in determining the strengths of the model in differentiating crops from weeds and an F1 score was useful as it balances both false positives and negatives. Further, to avoid or reduce such biases confusion matrices were produced that give a clear distinction of the model on class to class basis.

V. Results

In the results section, the performance of the Vision Transformer (ViT) model on crops, weeds, and background elements of the CWFID dataset is described in detail. Essentially, percentage accuracy, precision, recall and F1 scores were determined and more detailed analysis was done using the confusion matrix. The model was trained using 80 / 20 train-test split which helped evaluate the model on the new data it has never seen.

A. Accuracy Assessment

On the test set it was possible to obtain an overall accuracy of the ViT model equal to 89.4% showing that it can effectively distinguish crop, weed and background pixels. This high level of accuracy indicate that the ViT model is able to extract the unique features of each class even in the complicated agricultural environments where crops resemble weeds. The degree of accuracy shown in this paper proves that ViT model can be used in practical applications, specifically in the field of precision agriculture where precise identification of crops and weeds can lead to improvement in crop management and decrease in the amount of applied herbicide.

B. Class-Specific Performance

Class-specific precision, recall, and F1 scores were calculated to evaluate the model's effectiveness across different classes: crops, weeds, and background. These metrics are as follows and are summarised in Table III for easy comparison of the strengths and weaknesses of the model with respect to each class. Fig. 3 shows various models performance results.

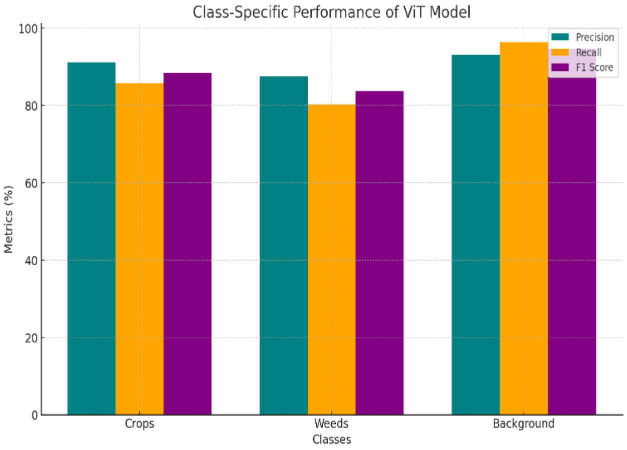

Fig. 3. Various models performance results.

TABLE III. Specific Performance Metrics of ViT Model

Class	Precision (%)	Recall (%)	F1 Score (%)
Crops	91.2	85.7	88.4
Weeds	87.5	80.3	83.7
Background	93.1	96.4	94.7

For the crop detection, the model obtained an accuracy of 91.2% and recall of 85.7% and thus an F1 score of 88.4%. This is, however, high, although there could be confusion with weeds particularly in the inter-row area. Varying results were achieved for the recognition of weeds, with 87.5% accuracy, 80.3% recall, and thus an F1 of 83.7%. The slightly lower recall for weeds shows that weed detection is more difficult especially for intra row weeds which are more similar in appearance to the crops. In the evaluation of the model for background region, the precision achieved was 93.1%, with a recall of 96.4% and F1 score 94.7%. This high performance on the background further enhances the performance of the model in differentiating the non-vegetation areas, thus minimizing chances of wrongly classifying crops as weeds.

C. Confusion Matrix Analysis

The confusion matrix extends the assessment of the model's classification correctness by showing where the errors were made. In Table IV the true positive, the false positive, and the false negative are shown for each class.

The confusion matrix (Fig. 4) also shows that the major misclassification problem was between the crop and weed classes where crop pixels amounted to 168 were misclassified as weeds while weed pixels of 128 were classified as crops. This pattern indicates that the yarn becomes problematic in distinguishing between crops and weeds mainly within areas of high plant density. This is especially problematic in intra-row spaces where weeds and crops may have similar architectures and reflectance properties hence compounding the challenge of modeling the two. On the other hand, the background class was accurately classified with few errors which actually shows that the model is good in separating vegetative from the non-vegetative land cover like the soiler bare ground.

TABLE IV. CONFUSION MATRIX FOR VIT MODEL PREDICTIONS ON TEST SET

	Predicted Crop	Predicted Weed	Predicted Background
Actual Crop	1,221	128	13
Actual Weed	168	1,345	72
Actual Background	7	12	1,249

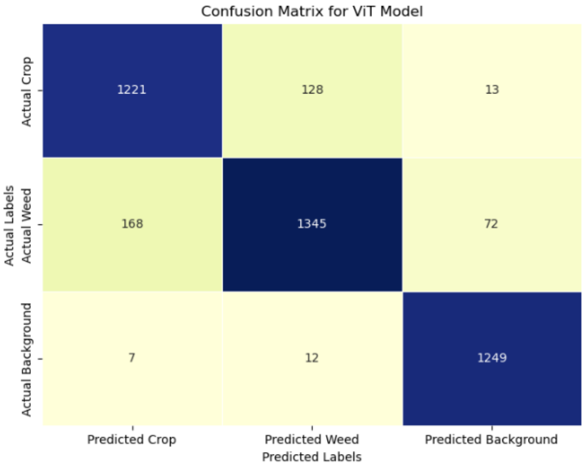

Fig. 4. Confusion matrix.

D. Comparison with Other Models

The ViT model was compared with traditional models such as U-Net, SegNet, and Fully Convolutional Network (FCN). Accuracy, precision, recall and F1 scores of all models are summarized in the Table V, where it can be concluded that the ViT model is more accurate. By comparing the proposed Vision Transformer (ViT) model with those of U-Net, SegNet, and Fully Convolutional Network (FCN), its higher accuracy has established it to be capable of handling the difficult environments within agriculture, especially the growth within the intra-row weed.

In such environments, where weeds are below or adjacent to crops and may be morphologically similar to crops, many of the CNN-based models are ineffective. This is due to the fact that, convolutional layers are inherently limited by its local receptive field, meaning that traditional model might be unable to capture those high-level, global features requiring the understanding of the whole image and its relationship to all other images, which in turn affects its accuracy in situations where high level of discriminative dissimilarities exists.

The self-attention mechanism of the ViT model has an advantage because it processes images in their entirety and identifies long-range spatial relations that may be neglected by CNN-based architectures. Such an approach is most beneficial for intra-row weed identification, in which local resemblance in texture and color between crops and weeds often leads to confusion in other models. This paper also shows that self-attention mechanism in ViT that allows the model to pay attention to relevant features in large regions of the images leads to better recall and precision, important for weed classification where precise distinction between crop and weed pixels is necessary (see Fig. 5, 6 and 7).

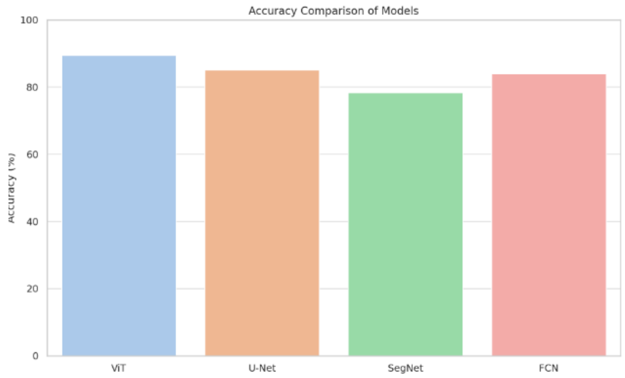

Fig. 5. Accuracy comparisions of models.

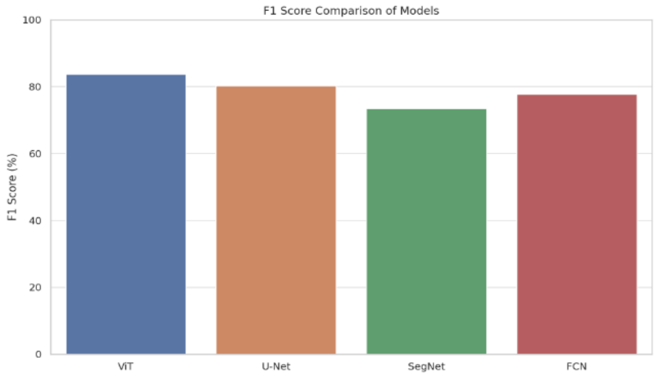

Fig. 6. F1-Score comparisions of models.

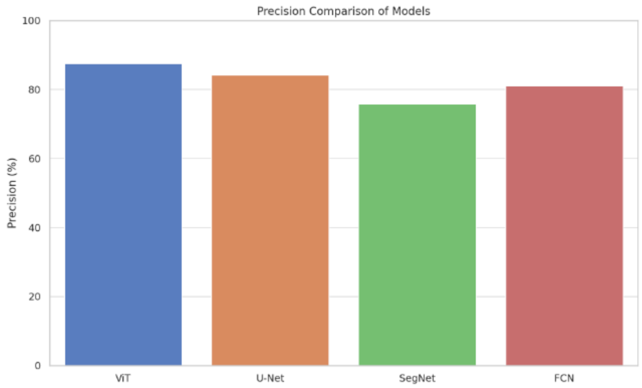

Fig. 7. Precision comparisions of models.

TABLE V. PERFORMANCE COMPARISON OF VIT MODEL WITH TRADITIONAL MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
ViT	89.4	87.5	80.3	83.7
U-Net	85.1	84.2	76.5	80.2
SegNet	78.3	75.8	71.4	73.5
FCN	83.9	81.1	74.8	77.8

Furthermore, the improved performance in ViT can again be attributed to how generalization is done properly to the field conditions. The high accuracy and precision in signifying weeds and crop images justify the flexibility of the proposed model in different lighting backgrounds, soil, and crop placement. This generalization is especially helpful when working with practical field applications of the model as weather, lighting, and growth stages may affect the models’ performance. A noteworthy comparison with U-Net further emphasizes ViT's advantage: that although U-Net also yields a good performance, the reliance on convolutional layers again hinders the ability to capture global context and therefore yields lower recall for weed detection. The results indicate that the trained U-Net is more sensitive to vagaries in closely planted weeds and crops, issues that are worsened by field conditions. However, these difficulties are not present in ViT’s design, which implies that potential agricultural applications of transformer-based models might be more scalable and versatile where extensive field analysis is required.

VI. CONCLUSION

The current paper shows that Vision Transformers (ViTs) can be used in precision agriculture for the detection of weeds in crop fields. This accuracy was established through pixel-level classification adopted from high-resolution UAV imagery as compared to traditional models such as U-Net and FCN where the ViT model obtained 89.4% accuracy. The high accuracy is an indication of ViT’s ability to establish dependencies and spatial arrangement in large agricultural scenes, which are hard for traditional CNNs to achieve. The study achieved the following goals of the research: The class imbalance was solved by applying a combination of two oversampling techniques which improved the classification results. Employing patch extraction and data augmentation enabled the ViT model to accurately distinguish crop, weed and background regions. The approach also showed robustness under different conditions improving the likelihood of its application in realistic agricultural settings.

This research goes a long way in the promotion of sustainable agriculture by providing a potential method for selective weed management that does not require much use of the weed controlling herbicide. The current study could be extended in the future by examining other environmental factors or using the model in other crop types, and different field conditions to assess the model’s universality. Finally, the model derived from ViT holds the potential to contribute toward precise, effective and sustainable farming.

This research also opens up possibilities for integration with AI-based tractors, enabling real-time weed detection and management directly in the field. Such applications could revolutionize automated precision agriculture, allowing for targeted weed control while minimizing herbicide usage. With further development, this approach could support the advancement of intelligent, autonomous farming machinery.

REFERENCES

[1] S. Nath, "A vision of precision agriculture: Balance between agricultural sustainability and environmental stewardship," Agronomy Journal, vol. 116, no. 3, pp. 1126-1143, 2024.

[2] L. De Bortoli, S. Marsi, F. Marinello, and P. Gallina, "Cost-efficient algorithm for autonomous cultivators: Implementing template matching with field digital twins for precision agriculture," Computers and Electronics in Agriculture, vol. 227, p. 109509, 2024.

[3] D. C. Brainard et al., "A survey of weed research priorities: key findings and future directions," Weed Science, vol. 71, no. 4, pp. 330-343, 2023.

[4] M. Vasileiou et al., "Transforming weed management in sustainable agriculture with artificial intelligence: A systematic literature review towards weed identification and deep learning," Crop Protection, p. 106522, 2023.

[5] F. Yeganehpoor, S. Z. Salmasi, G. Abedi, F. Samadiyan, and V. Beyginiya, "Effects of cover crops and weed management on corn yield," Journal of the Saudi Society of Agricultural Sciences, vol. 14, no. 2, pp. 178-181, 2015.

[6] P. Hatcher and B. Melander, "Combining physical, cultural and biological methods: prospects for integrated non - chemical weed management strategies," Weed research, vol. 43, no. 5, pp. 303-322, 2003.

[7] I. Bhakta, S. Phadikar, and K. Majumder, "State - of - the - art technologies in precision agriculture: a systematic review," Journal of the Science of Food and Agriculture, vol. 99, no. 11, pp. 4878-4888, 2019.

[8] N. Zhang, M. Wang, and N. Wang, "Precision agriculture—a worldwide overview," Computers and electronics in agriculture, vol. 36, no. 2-3, pp. 113-132, 2002.

[9] D. C. Tsouros, S. Bibi, and P. G. Sarigiannidis, "A review on UAV-based applications for precision agriculture," Information, vol. 10, no. 11, p. 349, 2019.

[10] A. Upadhyay et al., "Advances in ground robotic technologies for site-specific weed management in precision agriculture: A review," Computers and Electronics in Agriculture, vol. 225, p. 109363, 2024.

[11] A. H. Al-Badri et al., "Classification of weed using machine learning techniques: a review—challenges, current and future potential techniques," Journal of Plant Diseases and Protection, vol. 129, no. 4, pp. 745-768, 2022.

[12] F. D. Adhinata and R. Sumiharto, "A comprehensive survey on weed and crop classification using machine learning and deep learning," Artificial Intelligence in Agriculture, 2024.

[13] R. Reedha, E. Dericquebourg, R. Canals, and A. Hafiane, "Transformer neural network for weed and crop classification of high resolution UAV images," Remote Sensing, vol. 14, no. 3, p. 592, 2022.

[14] S. Sharma and M. Vardhan, "Self-attention vision transformer with transfer learning for efficient crops and weeds classification," in 2023 6th International Conference on Information Systems and Computer Networks (ISCON), 2023: IEEE, pp. 1-6.

[15] T. Burks, S. Shearer, and F. Payne, "Classification of weed species using color texture features and discriminant analysis," Transactions of the ASAE, vol. 43, no. 2, pp. 441-448, 2000.

[16] Z. Wu, Y. Chen, B. Zhao, X. Kang, and Y. Ding, "Review of weed detection methods based on computer vision," Sensors, vol. 21, no. 11, p. 3647, 2021.

[17] M. N. Khan and S. Anwar, "Robust weed recognition through color based image segmentation and convolution neural network based classification," in ASME International Mechanical Engineering Congress and Exposition, 2019, vol. 59414: American Society of Mechanical Engineers, p. V004T05A045.

[18] S. Lavania and P. S. Matey, "Novel method for weed classification in maize field using Otsu and PCA implementation," in 2015 IEEE International Conference on Computational Intelligence & Communication Technology, 2015: IEEE, pp. 534-537.

[19] H. Huang, Y. Lan, A. Yang, Y. Zhang, S. Wen, and J. Deng, "Deep learning versus Object-based Image Analysis (OBIA) in weed mapping of UAV imagery," International Journal of Remote Sensing, vol. 41, no. 9, pp. 3446-3479, 2020.

[20] M. H. M. Zaman, S. M. Mustaza, M. F. Ibrahim, M. A. Zulkifley, and M. M. Mustafa, "Weed classification based on statistical features from Gabor transform magnitude," in 2021 International Conference on Decision Aid Sciences and Application (DASA), 2021: IEEE, pp. 147-151.

[21] P. Bosilj, T. Duckett, and G. Cielniak, "Analysis of morphology-based features for classification of crop and weeds in precision agriculture," IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 2950-2956, 2018.

[22] P. Rasti, A. Ahmad, S. Samiei, E. Belin, and D. Rousseau, "Supervised image classification by scattering transform with application to weed detection in culture crops of high density," Remote Sensing, vol. 11, no. 3, p. 249, 2019.

[23] H. Okamoto, T. Murata, T. Kataoka, and S. I. HATA, "Plant classification for weed detection using hyperspectral imaging with wavelet analysis," Weed biology and Management, vol. 7, no. 1, pp. 31-37, 2007.

[24] C. Feng et al., "A combination of OBIA and random forest based on visible UAV remote sensing for accurately extracted information about weeds in areas with different weed densities in farmland," Remote Sensing, vol. 15, no. 19, p. 4696, 2023.

[25] C.-C. Andrea, B. B. M. Daniel, and J. B. J. Misael, "Precise weed and maize classification through convolutional neuronal networks," in 2017 IEEE Second zcuador Technical Chapters Meeting (ETCM), 2017: IEEE, pp. 1-6.

[26] M. Fawakherji, C. Potena, A. Pretto, D. D. Bloisi, and D. Nardi, "Multi-spectral image synthesis for crop/weed segmentation in precision farming," Robotics and Autonomous Systems, vol. 146, p. 103861, 2021.