

# Security Analysis and Robustness Assessment of Agricultural Deep Vision Software Against Targeted Adversarial Evasion Attacks in Crop-Weed Classification

Dr.J. Justina Michael<sup>1</sup>, Dr.H. Mary Shyni<sup>2\*</sup>, G. Praveen Kumar<sup>3</sup>, and Dr.U. Sakthivelu<sup>4</sup>

<sup>1</sup>Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India. justinaj@srmist.edu.in, <https://orcid.org/0000-0001-8072-3230>

<sup>2\*</sup>Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani, Chennai, Tamil Nadu, India. maryshyh@srmist.edu.in, <https://orcid.org/0000-0002-6386-5947>

<sup>3</sup>Department of Information Technology, St. Joseph's Institute of Technology, OMR, Semmencheri, Chennai, Tamil Nadu, India. praveenkumarg@stjosephstechnology.ac.in, <https://orcid.org/0009-0004-3668-7390>

<sup>4</sup>Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India. sakthiveluudayakumar@gmail.com, <https://orcid.org/0009-0004-4971-5338>

Received: October 25, 2025; Revised: December 02, 2025; Accepted: January 21, 2026; Published: February 27, 2026

## Abstract

This study aims to evaluate the robustness and security of deep vision-based crop–weed segmentation models against targeted adversarial evasion attacks and to assess the effectiveness of defense mechanisms for improving reliability in precision agriculture systems. Experiments were conducted using the SorghumWeedDataset\_Segmentation dataset containing 5,555 annotated segments from 252 high-resolution RGB images of sorghum, grasses, and broad-leaf weeds. Convolutional neural networks were trained using state-of-the-art methods, and they were used to perform pixel-based segmentation and classification. Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) were used to produce adversarial examples at different magnitudes of perturbations ( $\epsilon$ ). The measures of model performance were accuracy, precision, recall, F1-score, Intersection over Union (IoU), Attack Success Rate (ASR), and a robustness measure. Adversarial training was introduced in order to increase resilience. The model had 94.8% classification and a 0.887 mean IoU under clean conditions. Nevertheless, it deteriorated considerably during adversarial perturbations. However, performance degraded significantly under adversarial perturbations. At  $\epsilon = 0.03$ , FGSM reduced accuracy to 68.4% and mean IoU to 0.612 (ASR: 31.6%), while PGD caused a sharper decline to 54.7% accuracy and 0.482 mean IoU (ASR: 45.3%). After adversarial training, adversarial accuracy improved from 68.4% to 81.2%, and mean IoU increased from 0.612 to 0.743, with only a marginal reduction in clean accuracy (94.8% to 92.6%). The results indicate that deep agricultural vision models are very susceptible to gradient-based adversarial attacks, especially iterative ones, such as PGD. The Introduction of adversarial training makes it much more robust and trades off a little clean performance. Robustness

---

*Journal of Internet Services and Information Security (JISIS)*, volume: 16, number: 1 (February-2026), pp. 849-863. DOI: 10.58346/JISIS.2026.11.049

\*Corresponding author: Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani, Chennai, Tamil Nadu, India.

evaluation is therefore essential for the secure deployment of AI-driven weed management systems in real-world agricultural environments.

**Keywords:** Adversarial Attacks, Deep Learning, Precision Agriculture, Robustness, Semantic Segmentation.

## 1 Introduction

The fact is that the drastic development of deep learning has transformed agriculture, especially the implementation of computer vision technology in crop monitoring and weed management. Deep neural networks (DNNs) in automated crop-weed classification have made possible precision agriculture methods that have led to the reduction of labor expenses, enhanced resource management, and enhanced productivity. These systems can be used to accurately differentiate between crops and weeds, which helps in targeted application of herbicides, reducing the use of chemicals and environmental effects. Although DNNs achieve a high-performance rate when used in a controlled environment, they also inherently suffer adversarial attacks that are carefully constructed perturbations to trick the input images into misclassifications, yet are not visible to human eyes (Gao et al., 2024). Such misclassifications may lead to incorrect weed elimination, damage to crops, and economic losses in agricultural settings, and therefore, the immediate necessity to study the security and strength of these systems in a real-life application (Yazdinejad et al., 2021).

The main aim of this research is to evaluate the susceptibility of deep vision-based agricultural software to adversarial evasion attacks and also to test methods of enhancing model resistance. The study seeks to establish how much adversarial manipulation can impair classification by methodically examining how the manipulation of crop and weed images affects their classification (Qu & Su, 2024; San & Kakani, 2025). Moreover, the research examines defense strategies, such as adversarial training and perturbation-resilient model adaptations, in order to assess the success of all of them in correcting attacks on their systems (Mowla, 2021).

Adversarial attacks and defense strategies are well studied in the general computer vision domains; however, there is limited research done in the specific area of agriculture. The majority of current crop-weed classification models are tested using ideal or clean datasets, which ignore the risks that could be generated by adversarial interference in the field. Also, the systematic evaluation of both targeted and untargeted attack conditions and their effects on the model reliability is lacking. The identified research gap highlights the necessity of special research that will span the areas of agricultural deep learning and security analysis to offer real-life insights into their application (Zhao & Wang, 2025).

In this hypothesis, although deep vision models are highly accurate on traditional datasets, agricultural deep vision models are vulnerable to adversarial perturbations that may cause major misclassification between crops and weeds. Strategies that are more robust, including adversarial training or detectors, can enhance model robustness and guarantee it operates reliably in adversarial environments (Pai et al., 2024).

The main contributions made in this work are four. It first gives the complete security analysis of state-of-the-art models of crop-weed classification to targeted adversarial attacks. Second, it measures model weaknesses at different levels of attack strengths and conditions. Third, it suggests and analyzes the enhancement of robustness strategies modified to suit agricultural purposes. Lastly, the research results provide viable recommendations on how to implement secure and reliable deep learning systems to support safer and more efficient practices of crop management in the field of precision agriculture.

This paper is divided into six key sections. The Introduction provides the motivation, research objectives, hypothesis, and main contributions in connection to the adversarial robustness in agricultural deep learning. The Literature Survey reviews prior work on crop–weed classification and adversarial security. The Materials and Methods section describes the SorghumWeedDataset\_Segmentation, preprocessing, deep vision model, adversarial attack generation (FGSM and PGD), metrics of robustness, and the experiment. Results contain objective and antagonistic performance appraisals and are quantitative analyses. The Discussion makes sense of significant findings and limitations. Lastly, there is the Conclusion, which lists the contributions and future research directions of secure precision agriculture deployment.

## 2 Literature Survey

The use of artificial intelligence (AI) and deep learning in agriculture has contributed greatly to precision farming, especially in crop monitoring, weed detection, and autonomous agricultural equipment. Recent surveys have noted that deep learning-based crop-weed recognition frameworks have demonstrated high accuracy within a controlled setting, allowing intelligent equipment to make real-time discoveries and spray onto specific regions (Qu & Su, 2024; Jia et al., 2025). Convolutional neural networks (CNNs), U-Net, and transformer-based models are commonly used in semantic and instance segmentation of agricultural images. Likewise, (Rizvi et al., 2024) indicates the revolutionary potential of machine learning and deep learning to enhance the quality of crops as well as the efficiency of weeding machines at a higher level (Garcia-Oliveira et al., 2026; Tarek et al., 2023).

Although all this is happening, the topic of security and robustness issues in agricultural AI is under-researched. There are new security risks that are aimed at agricultural AIs, such as adversarial manipulation, data poisoning, and vulnerabilities in the system of autonomous farming machinery (Zhang et al., 2025). Detection of cyber-physical attacks in smart farming systems with emphasis on safe model implementation. IoT-based agriculture presents new attack venues, such as sensor-spoofing and vulnerabilities of the communication layers (Naseer et al., 2024; Enoch et al., 2026).

Although significant advancements have been achieved in deep learning-based weed detection, the majority of research tests deep learning models in clean datasets, and little focus is on the adversarial robustness of the models. Accept environmental variations without paying much attention to intentional adversarial perturbations (Gill et al., 2022; Wu et al., 2025). Likewise, smart precision weeding and AI-based cooperative control systems have applications that are based on efficiency and automation, and not security (Divyanth et al., 2022; Xu et al., 2025).

The increased use of AI in agri-food engineering and plant breeding further increases the motivation to have secure and trustworthy AI systems (Prakash & Pravinesh, 2024; Sun et al., 2025). Nevertheless, there are no systematic analyses of adversarial evasion attacks in crop- weed recognition. This thus clearly indicates that there is a major gap between the high-performance agricultural vision systems and the security robustness, and thus the need for precise adversarial vulnerability assessment and model design that is defense-aware is indeed imperative in precision agriculture.

### 3 Materials and Methods

#### Dataset Description and Preprocessing

This study utilizes the Sorghum Weed Dataset Segmentation, a publicly available crop-weed research dataset aimed at addressing real-time weed management challenges through computer vision applications. The dataset consists of 5555 hand-labeled data segments of 252 high-resolution (RGB) images, which include sorghum plants (Class 0), grasses (Class 1), and broad-leaf weeds (Class 2). The data was collected in the period of April and May 2023 in the Sri Ramaswamy Memorial (SRM) Care Farm, in the Chengalpattu district, Tamil Nadu, India, and under varying weather and lighting conditions in order to generalize (Michael, 2023).

The images were shot with a Canon EOS 80D DSLR camera with a 22.3 mm x 14.9 mm CMOS sensor in a resolution of 6000 x 4000 pixels (~13 MB). The dataset contains manually pixel-wise annotated ground truth files in JSON, CSV, and COCO formats, which may be used in object detection, instance segmentation, and semantic segmentation. The dataset is divided into training, validation, and testing sets in an 8:1:1 ratio, and the samples are well balanced. The preprocessing steps include the resizing of images to standard resolution, pixel intensity normalization, and augmentation of images by random rotation, flips, and the intensity of the light to improve generalization in the models.

#### Deep Vision Model Architecture

Crop-weed classification and segmentation are achieved by using the state-of-the-art convolutional neural networks (CNNs), such as ResNet and EfficientNet series. The input image  $X \in \mathbb{R}^{H \times W \times C}$  passes through multiple convolutional layers  $F_i$  with learnable weights  $W_i$  and biases  $b_i$ , producing feature maps:

$$F_i = f(W_i * X + b_i) \quad (1)$$

In equation (1)  $*$  represents convolution and  $f$  is the activation function (ReLU). A softmax layer generates the class probabilities for sorghum, grasses, and broad-leaf weeds:

$$\hat{y} = \text{softmax}(Z) = \frac{e^{Z_j}}{\sum_{k=1}^K e^{Z_k}} \quad (2)$$

In equation (2),  $Z$  denotes output logits and  $K = 3$  is the number of classes. For segmentation tasks, the output is a per-pixel class prediction corresponding to the annotated ground truth masks.

#### Adversarial Attack Generation

Targeted adversarial attacks are generated to evaluate model robustness. The Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are used to craft perturbations  $\delta$  for each input image  $X$ , creating adversarial samples  $X'$ :

$$X' = X + \epsilon \cdot \text{sign}(\nabla_X \mathcal{L}(\theta, X, y)) \quad (3)$$

In equation (3)  $\epsilon$  is the perturbation magnitude and  $\mathcal{L}$  is the cross-entropy loss. PGD iteratively applies small perturbations while keeping them within a predefined  $\ell_\infty$  norm bound.

## Robustness Assessment

Model robustness is quantified using classification and segmentation accuracy, adversarial success rate, and a robustness metric  $R$  shown as equation (4):

$$R = 1 - \frac{\# \text{ successful adversarial misclassifications}}{\text{Total adversarial samples}}, \quad (4)$$

Experiments are conducted across different perturbation magnitudes  $\epsilon$ , and defense strategies, including adversarial training and input preprocessing, are evaluated for their effectiveness in improving robustness.

**Algorithm 1:** Robustness Assessment of Crop-Weed Segmentation Models Using SorghumWeedDataset\_Segmentation

### Input:

- Labeled dataset  $D = \{(X_i, Y_i)\}_{i=1}^N$  from SorghumWeedDataset\_Segmentation, where  $X_i$  is the input RGB image and  $Y_i$  is the pixel-wise annotated mask
- Deep vision model  $M(\theta)$  for segmentation and classification
- Perturbation magnitude  $\epsilon$
- Number of attack iterations  $T$  (for iterative attacks like PGD)
- Loss function  $\mathcal{L}$  (cross-entropy for classification, pixel-wise segmentation loss for masks)

### Output:

- Clean segmentation and classification accuracy  $A_{\text{clean}}$
- Adversarial segmentation and classification accuracy  $A_{\text{adv}}$
- Robustness metric  $R$

### Step 1: Dataset Preparation

1. Load SorghumWeedDataset\_Segmentation images and pixel-wise annotations in JSON, CSV, or COCO format.
2. Preprocess images: resize to uniform resolution, normalize pixel values.
3. Apply data augmentation: random rotations, flips, and brightness variations.
4. Split the dataset into training, validation, and testing sets with an 8:1:1 ratio.

### Step 2: Model Training

1. Initialize deep vision segmentation model  $M(\theta)$  (e.g., ResNet, EfficientNet, U-Net variant).
2. For each batch in the training set:
  - a. Forward pass: compute predicted segmentation masks  $\hat{Y} = M(X)$
  - b. Compute pixel-wise loss  $\mathcal{L}(\hat{Y}, Y)$
  - c. Backpropagate gradients and update model parameters  $\theta$
3. Validate model performance on the validation set and select the best-performing model.

### Step 3: Adversarial Example Generation

#### 3a: FGSM (Fast Gradient Sign Method)

For each test image  $X$  and corresponding mask  $Y$ :

$$X' = X + \epsilon \cdot \text{sign}(\nabla_X \mathcal{L}(M(\theta), X, Y))$$

#### 3b: PGD (Projected Gradient Descent, iterative)

1. Initialize  $X'_0 = X$
2. For  $t = 1$  to  $T$ :

$$X'_t = \text{Clip}_{X,\epsilon}(X'_{t-1} + \alpha \cdot \text{sign}(\nabla_{X'_{t-1}} \mathcal{L}(M(\theta), X'_{t-1}, Y)))$$

where  $\alpha$  is the step size and  $\text{Clip}_{X,\epsilon}$  ensures the adversarial image remains within  $\ell_\infty$ -norm bounds.

#### Step 4: Robustness Evaluation

1. Predict segmentation masks on clean test images:  $\hat{Y}_{\text{clean}} = M(X)$
2. Predict segmentation masks on adversarial images:  $\hat{Y}_{\text{adv}} = M(X')$
3. Compute evaluation metrics:
  - Pixel-wise IoU, precision, recall, and F1-score for segmentation
  - Classification accuracy of sorghum, grasses, and broad-leaf weeds
  - Robustness metric:

$$R = 1 - \frac{\# \text{ successful adversarial misclassifications}}{N}$$

#### Step 5: Defense and Mitigation

1. Apply adversarial training or input preprocessing techniques.
2. Repeat Steps 2–4 to measure improvements in  $A_{\text{adv}}$  and R.

The proposed algorithm 1 measures the strength of the deep learning-based crop-weed segmentation framework with the help of the SorghumWeedDataset\_Segmentation. The data is preprocessed and divided into training, validation, and test data, and a segmentation model is trained to classify at the pixel level sorghum, grasses, and broad leaf weeds. Adversarial examples are produced on FGSM and PGD attacks after the baseline assessment on clean images to add controlled perturbations. The performance of the model on adversarial inputs is then evaluated in terms of IOU, precision, recall, F1-score, and accuracy. A robustness measure can be used to measure vulnerability, allowing model security to be analyzed systematically and enhanced against evasion attacks targeted.

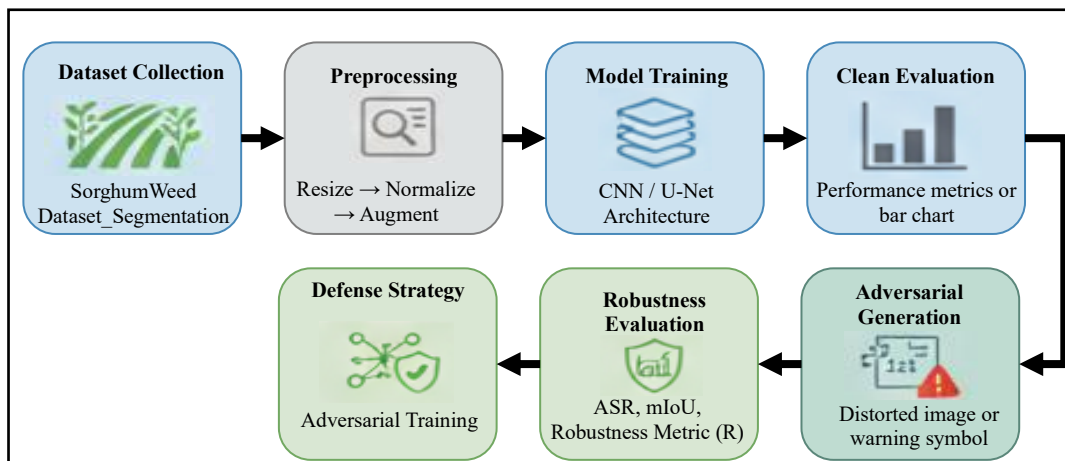


Figure 1: Study workflow for robustness evaluation framework

The full pipeline of the experiment, which involves the method of testing the robustness of models in crop-weed segmentation, is presented in Figure 1. The workflow starts with the collection of a dataset of the SorghumWeedDataset\_Segmentation, and then there is the process of preprocessing of the image, such as image resizing, normalization, and augmentation. CNN/U-Net model is then trained and tested

in clean conditions. FGSM and PGD attacks are necessary to create adversarial samples that measure model vulnerability. Lastly, the metrics of robustness evaluation and adversarial training are used to enhance the resilience of a model and achieve the provision of safe deployment in precision agriculture systems.

## Experimental Setup

All the models are written in Python on TensorFlow and PyTorch. The dataset is divided into train, validation, and test sets in the proportion of 8:1:1. Some of the evaluation measures include precision, recall, F1-score, pixel-wise IoU, and adversarial robustness. Model performance is statistically analyzed when the model is operated in clean conditions and adversarial conditions, which gives insight into the reliability of the model to be used in the field.

## 4 Results

### Performance on Clean Dataset

The trained deep vision segmentation model was first evaluated on the clean test subset of the SorghumWeedDataset\_Segmentation dataset. Performance was measured using classification accuracy, precision, recall, F1-score, and Intersection over Union (IoU).

**Classification accuracy is defined as:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

In equation (5),  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  represent true positives, true negatives, false positives, and false negatives, respectively.

Precision measures the correctness of positive predictions shown as equation (6):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

Recall quantifies the model's ability to detect true positives, shown as equation (7):

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

The F1-score, representing the harmonic mean of precision and recall, is calculated as equation (8):

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

For segmentation performance, Intersection over Union (IoU) is used as equation (9):

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (9)$$

The mean IoU (mIoU) is obtained by averaging IoU across all classes, shown as equation (10):

$$\text{mIoU} = \frac{1}{K} \sum_{k=1}^K \text{IoU}_k \quad (10)$$

where  $K = 3$  represents sorghum, grasses, and broad-leaf weeds.

The general classification accuracy on clean test data was 94.8%, and the mean Intersection over Union (mIoU) of the three classes was 0.887. Sorghum was most accurately segmented as it had a morphological entity that was well-structured, though the grasses and broad-leaf weeds had a slight degradation of the IoU values because of the overlapping features and complex backgrounds.

Table 1: Performance metrics on clean test data

Class	Precision	Recall	F1-Score	IoU
Sorghum (Class 0)	0.96	0.95	0.955	0.91
Grasses (Class 1)	0.93	0.92	0.925	0.87
Broad-leaf (Class 2)	0.94	0.93	0.935	0.88
Overall / Mean	0.94	0.93	0.938	0.887

Table 1 presents the performance of the clean test data in class segmentation and the general segmentation. The most accurate (0.96), recalled (0.95), F1-score (0.955), and the highest IoU (0.91) was Sorghum (Class 0), as it is possible to recognize it because of its specifics. A reduction was seen when it came to the grasses (Class 1) and broad-leaf weeds (Class 2), whose IoU values were 0.87 and 0.88, respectively, presumably owing to shared visual characteristics and complexity. The average values of precision (0.94), recall (0.93), F1-score (0.938), and mIoU (0.887) demonstrate effective and trustworthy results of segmentation in clean and non-adversarial conditions.

### Impact of FGSM Adversarial Attacks on crop-weed

In order to measure robustness, adversarial samples were perturbed with the Fast Gradient Sign Method (FGSM) with increasing perturbation magnitude ( $\epsilon$ ). Any little perturbation greatly lowered model performance.

The classification accuracy also dropped comparatively at 86.5% at  $\epsilon = 0.01$ . However,  $\epsilon = 0.03$  resulted in a much lower accuracy of 68.4% and an average IoU of 0.612. There was a significant increase in misclassification of grasses and sorghum, implying that the model was highly dependent on fine-grained texture patterns that were prone to perturbations.

Table 2: Model performance under FGSM attack

Perturbation ( $\epsilon$ )	Accuracy	Mean IoU	Attack Success Rate
0.00 (Clean)	94.8%	0.887	0.0%
0.01	86.5%	0.791	13.5%
0.02	76.9%	0.702	23.1%
0.03	68.4%	0.612	31.6%

Table 2 shows the model performance with the increasing FGSM perturbation strengths ( $\epsilon$ ). The model had high accuracy and the mean IoU under clean conditions ( $\epsilon = 0.00$ ). However, with  $\epsilon$ , the accuracy and the mean IoU decreased gradually, whereas the Attack Success Rate (ASR) increased significantly. With  $\epsilon = 0.03$ , the accuracy reduced to 68.4, and the mean IOU to 0.612, and ASR went up to 31.6%. The findings indicate that minor adversarial perturbation has a considerable negative impact on segmentation reliability and exposes a system to misclassification.

### Impact of PGD Adversarial Attacks

Being an iterative and stronger attack, Projected Gradient Descent (PGD) led to greater performance degradation. The classification accuracy at  $\epsilon = 0.03$  with 10 iterations decreased to 54.7%, and the mean IoU to 0.482.

The segmentation masks showed visible distortion, particularly along object boundaries. Sorghum plants were frequently misclassified as grasses, and smaller broad-leaf weeds were incorrectly merged into background regions.

Table 3: Model performance under PGD attack

Perturbation ( $\epsilon$ )	Accuracy	Mean IoU	Attack Success Rate
0.01	78.2%	0.689	21.8%
0.02	63.5%	0.553	36.5%
0.03	54.7%	0.482	45.3%

Table 3 summarizes the model performance in PGD adversarial attacks at the strength of perturbation ( $\epsilon$ ) incrementally. Right after increasing the epsilon to 0.01 to 0.03, the classification rate declines rapidly, as 78.2% to 54.7%, and the mean IoU declines from 0.689 to 0.482. At the same time, the Attack Success Rate (ASR) also increases greatly (21.8% to 45.3%). The iterative PGD approach produces worse segmentation performance and object boundary demarcation, indicating the increased susceptibility of the model to more robust, multi-step adversarial perturbations, compared to single-step attacks (Figure 2).

Figure 2(A) presents the reduction in mean IoU with increasing perturbation magnitude. The segmentation performance deteriorates progressively under adversarial conditions, with PGD producing greater boundary distortion and class confusion than FGSM, demonstrating its higher impact on pixel-wise segmentation quality. Figure 2(B) shows a downward trend in the accuracy of the classification (%) with the ratio of the adversarial perturbation magnitude ( $\epsilon$ ). The findings indicate that although FGSM and PGD both substantially drop the accuracy, the iterative PGD attack leads to a larger drop than the FGSM, as the former performs better in the attack.

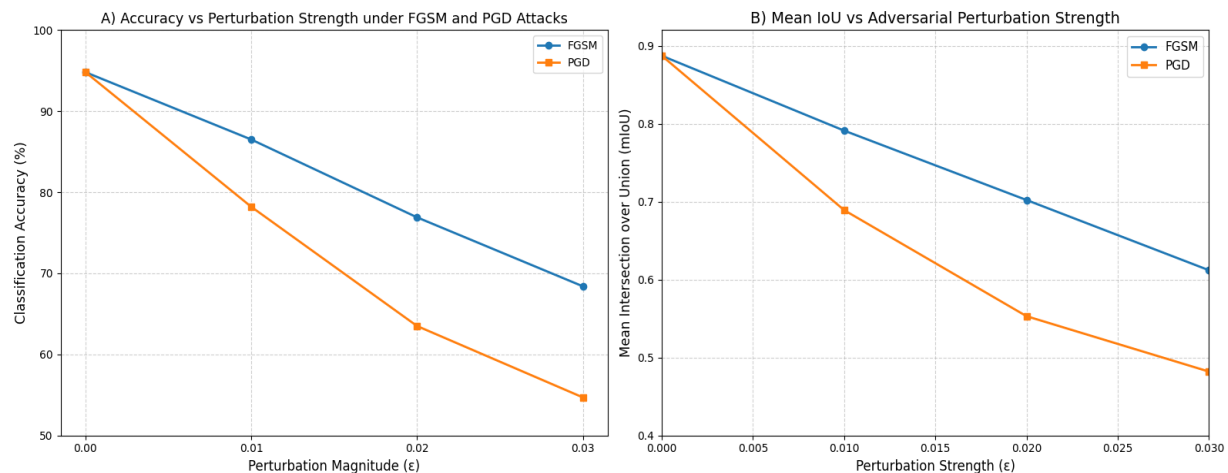


Figure 2: Robustness evaluation under adversarial perturbations. (A) classification accuracy vs. perturbation strength ( $\epsilon$ ) under FGSM and PGD Attacks. (B) mean intersection over union (mIoU) vs. perturbation strength ( $\epsilon$ )

### Robustness Metric Evaluation

Adversarial training was also included to strengthen the training by increasing the training data with adversarial samples generated by FGSM. The defense mechanism improved adversarial accuracy considerably while maintaining competitive clean performance.

After adversarial training, clean accuracy slightly reduced to 92.6%, but adversarial accuracy under FGSM ( $\epsilon = 0.03$ ) improved from 68.4% to 81.2%. Mean IoU also increased from 0.612 to 0.743 under the same attack condition.

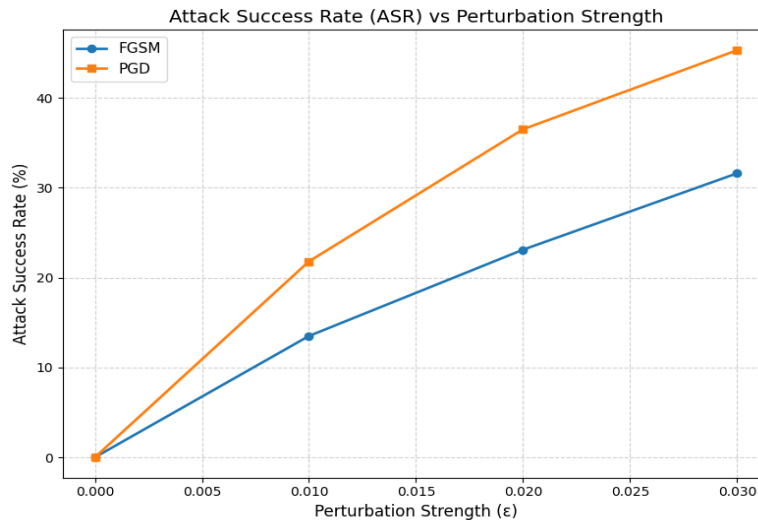


Figure 3: Attack Success Rate (ASR) under FGSM and PGD attacks across varying perturbation strengths ( $\epsilon$ )

Figure 3 shows the FGSM and PGD adversarial attack Success Rate (ASR) of various perturbation magnitudes ( $\epsilon$ ). The ASR is increasing steadily with  $\epsilon$  of both approaches, and this shows that there are higher misclassification rates in situations with more powerful perturbations. The iterative PGD attack has been shown to be much stronger against the single-step FGSM attack at all values of  $\epsilon$ , with its stronger adversarial influence, indicating that the model is more susceptible to attacks by adversarial gradients at an iterative scale.

Table 4: Performance before and after adversarial training ( $\epsilon = 0.03$ )

Model Type	Clean Accuracy	Adversarial Accuracy	Mean IoU (Adv)
Standard Model	94.8%	68.4%	0.612
Adversarially Trained	92.6%	81.2%	0.743

Table 4 contrasts the performance of the standard model and the adversarial trained model with  $\epsilon = 0.03$  perturbation. Although the standard model has a high clean accuracy (94.8%), their adversarial accuracy is much less (68.4%), with a mean IoU of 0.612. By contrast, the adversarial trained model has competitive clean accuracy (92.6%) with a significant increase in adversarial accuracy (81.2%) and mean IoU (0.743). These findings prove that adversarial training is very useful in improving the resilience to perturbations at a relatively small cost to clean performance.

Figure 4 compares the accuracy in classification in clean conditions and the most powerful attacks by adversaries ( $\epsilon = 0.03$ ). Although the standard model is highly performative on clean data, its accuracy reduces drastically in the case of FGSM and PGD attacks. Conversely, the adversarially trained model has significantly greater robustness (81.2% with FGSM), indicating that adversarial training is effective in reducing performance without any competitive clean accuracy.

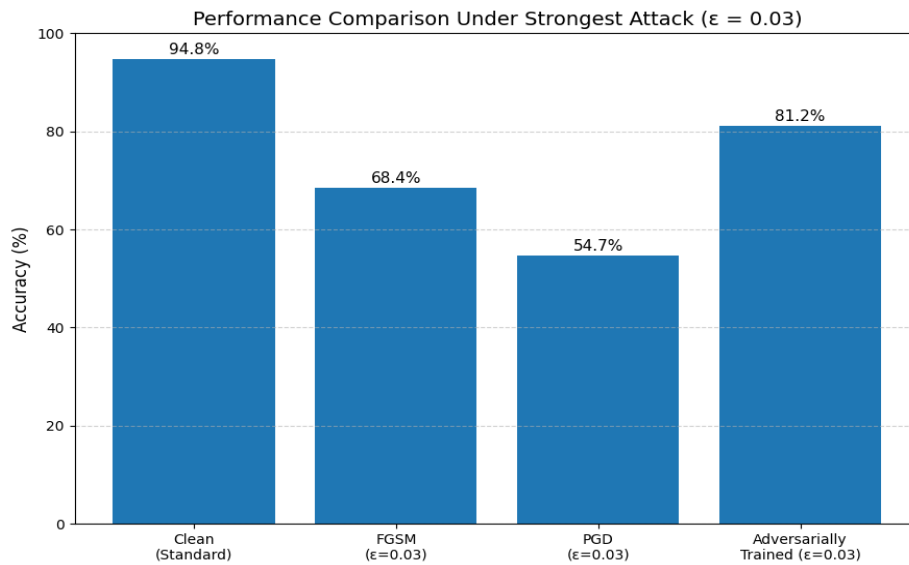


Figure 4: Performance comparison between standard and adversarially trained models under the strongest attack settings ( $\epsilon = 0.03$ )

### Overall Interpretation

The experimental findings prove that the segmentation model works very well in a clean environment, but it is also very prone to adversarial evasion attacks. FGSM induces moderate degradation, but PGD has a significant effect on the accuracy of segmentation and outlining object boundaries. The use of adversarial training significantly increases robustness, and clean accuracy trade-offs are minimal. These results highlight the fact that robustness assessment is critical prior to using deep vision to manage weeds in real-time in farm settings.

### Discussion

The findings show the deep vision segmentation model can be characterized by high performance in clean and non-adversarial situations, but is significantly susceptible to adversarial perturbations. The model showed high classification and segmentation qualities under clean evaluation, which means that the model learned features effectively in the differentiation of sorghum and weed. Nevertheless, FGSM and PGD attacks hugely deteriorated as the strength of perturbation ( $\epsilon$ ) was elevated. The PGD attack repeatedly generated more serious degradation compared to the one-step FGSM attack, showing how gradient-based vision systems are vulnerable to more powerful adversarial optimization. Significantly, the addition of adversarial training provided significant gains in robustness and did not diminish the competitive clean performance, which validates the success of defense-based training methods in agricultural vision tasks.

The model had a classification accuracy of 94.8%, and a mean IoU of 0.887 on clean data, and Sorghum had a high IoU of 0.91. FGSM attack with 0.03 also reduced accuracy to 68.4%, mean IoU to 0.612, and Attack Success Rate (ASR) rose to 31.6%. The PGD attack led to a greater degradation where the accuracy decreased to 54.7%, the mean IoU was 0.482, and ASR rose to 45.3% with  $\epsilon = 0.03$ . After adversarial training, adversarial accuracy improved from 68.4% to 81.2%, and mean IoU increased from 0.612 to 0.743, while clean accuracy showed only a minor reduction from 94.8% to 92.6%. These

numerical trends confirm that robustness enhancement is achievable with limited trade-offs in baseline performance.

Despite promising robustness improvements, several limitations remain. First, adversarial evaluation was restricted to FGSM and PGD attacks; other adaptive or black-box attacks may produce different vulnerability patterns. Second, there was adversarial training on FGSM-generated samples only, which might exaggerate generalization to more challenging iterative or unseen attack strategies. Third, the robustness evaluation was conducted in controlled experimental conditions on a particular dataset and in a three-class segmentation model, which might not reasonably capture the variability of the real world in agriculture, including variations in illumination, occlusions, and sensor noise. Lastly, the computational cost and the viability of real-time deployment through the defense mechanisms of adversarial defense were not assessed in detail. The work in the future should include various attack models, multi-defense strategies, and field-level validation to have secure deployment in the precision agriculture systems.

## 5 Conclusion

In this paper, the strength of a deep vision model to segment sorghum and weed within adversarial perturbation was assessed, and it was shown that robustness-aware training is significant in agricultural AI systems. The model demonstrated great success in clean conditions but failed tremendously when subjected to gradient-based adversarial attacks. The results validate that adversarial perturbations have the potential to significantly affect the stability of segmentation and boundary delineation, and can have implications for real-time decision-making in precision agriculture. The sharp idea of including adversarial training provided much higher robustness with a competitive clean output, which underlines its practical applicability in ensuring safe deployment. In clean conditions, the model performed well with a classification accuracy of 94.8% and a mean IoU of 0.887, showing good segmentation. But the adversarial evaluation showed that it is very vulnerable. FGSM at  $\epsilon = 0.03$  reduced accuracy to 68.4%, and mean IoU to 0.612, and PGD decreased accuracy to 54.7% with a mean IoU of 0.482 at 45.3% Attack Success Rate. These results confirm the fact that iterative attacks significantly undermine the reliability of segmentation. Adversarial training improved robustness considerably: adversarial accuracy increased from 68.4% to 81.2%, and mean IoU improved from 0.612 to 0.743, with only a marginal reduction in clean accuracy (94.8% to 92.6%). Future directions should be concerned with resistance to adaptive and black-box attacks, multi-step adversarial training approaches, and experimenting with hybrid defense methods that combine input preprocessing and model regularization. Additionally, real-time field validation under diverse environmental conditions is necessary to ensure secure deployment in precision agriculture systems. The added inclusion of lightweight and computationally efficient defense models will also increase the practical use for real-time weed management.

## References

- [1] Divyanth, L. G., Guru, D. S., Soni, P., Machavaram, R., Nadimi, M., & Paliwal, J. (2022). Image-to-image translation-based data augmentation for improving crop/weed classification models for precision agriculture applications. *Algorithms*, 15(11), 401. <https://doi.org/10.3390/a15110401>
- [2] Enoch, S. Y., Jibril, M. L., Malgwi, Y. M., & Wajiga, G. M. (2026). A comprehensive survey of applications, techniques, threats, and security countermeasures in smart farming. *Franklin Open*, 100490. <https://doi.org/10.1016/j.fraope.2026.100490>

- [3] Gao, Y., Camtepe, S. A., Sultan, N. H., Bui, H. T., Mahboubi, A., Aboutorab, H., ... & Singh, D. (2024). Security threats to agricultural artificial intelligence: Position and perspective. *Computers and Electronics in Agriculture*, 227, 109557. <https://doi.org/10.1016/j.compag.2024.109557>
- [4] Garcia-Oliveira, A. L., Dwivedi, S. L., Chander, S., Nelimor, C., Abd El Moneim, D., & Ortiz, R. O. (2026). Breeding Smarter: Artificial Intelligence and Machine Learning Tools in Modern Breeding A Review. *Agronomy*, 16(1), 137. <https://doi.org/10.3390/agronomy16010137>
- [5] Gill, T., Gill, S. K., Saini, D. K., Chopra, Y., de Koff, J. P., & Sandhu, K. S. (2022). A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. *Phenomics*, 2(3), 156-183. <https://doi.org/10.1007/s43657-022-00048-z>
- [6] Jia, H., Chen, W., Su, Z., Sun, Y., Qian, Z., & Huang, L. (2025). AI-Driven Cooperative Control for Autonomous Tractors and Implements: A Comprehensive Review. *AgriEngineering*, 7(11), 394. <https://doi.org/10.3390/agriengineering7110394>
- [7] Michael, J. (2023). *SorghumWeedDataset\_Segmentation* [Dataset]. <https://doi.org/10.17632/y9bmtf4xmr.1>
- [8] Mowla, M. N. (2021). *Weed detection and classification using deep learning* [Master's thesis, Adana Science and Technology University]. <https://doi.org/10.13140/rg.2.2.13329.71523>
- [9] Naseer, A., Shmoon, M., Shakeel, T., Rehman, S. U., Ahmad, A., & Gruhn, V. (2024). A systematic literature review of the IoT in agriculture—global adoption, innovations, security, and privacy challenges. *Ieee Access*, 12, 60986-61021. <https://doi.org/10.1109/access.2024.3394617>
- [10] Pai, D. G., Kamath, R., & Balachandra, M. (2024). Deep learning techniques for weed detection in agricultural environments: A comprehensive review. *IEEE Access*, 12, 113193-113214. <https://doi.org/10.1109/access.2024.3418454>
- [11] Prakash, L. R., & Pravinesh, H. (2024, November). A Novel Approach for Classification of Crops and Weeds using Deep Learning. In *2024 International Conference on Electronic Systems and Intelligent Computing (ICESIC)* (pp. 284-288). IEEE. <https://doi.org/10.1109/icesic61777.2024.10846080>
- [12] Qu, H. R., & Su, W. H. (2024). Deep learning-based weed-crop recognition for smart agricultural equipment: A review. *Agronomy*, 14(2), 363. <https://doi.org/10.3390/agronomy14020363>
- [13] Rizvi, S. M. H., Naseer, A., Rehman, S. U., Akram, S., & Gruhn, V. (2024). Revolutionizing agriculture: Machine and deep learning solutions for enhanced crop quality and weed control. *IEEE Access*, 12, 11865-11878. <https://doi.org/10.1109/access.2024.3355017>
- [14] San, C. T., & Kakani, V. (2025). Smart precision weeding in agriculture using 5ir technologies. *Electronics*, 14(13), 2517. <https://doi.org/10.3390/electronics14132517>
- [15] Sun, H., Chu, H. Q., Qin, Y. M., Hu, P., & Wang, R. F. (2025). Empowering Smart Soybean Farming with Deep Learning: Progress, Challenges, and Future Perspectives. *Agronomy*, 15(8), 1831. <https://doi.org/10.3390/agronomy15081831>
- [16] Tarek, Z., Elhoseny, M., Alghamdi, M. I., & El-Hasnony, I. M. (2023). Leveraging three-tier deep learning model for environmental cleaner plants production. *Scientific Reports*, 13(1), 19499. <https://doi.org/10.1038/s41598-023-43465-4>
- [17] Wu, K., Ji, Z., Wang, H., Shao, X., Li, H., Zhang, W., ... & Bao, X. (2025). A comprehensive review of AI methods in agri-food engineering: Applications, challenges, and future directions. *Electronics*, 14(20), 3994. <https://doi.org/10.3390/electronics14203994>
- [18] Xu, B., Werle, R., Chudzik, G., & Zhang, Z. (2025). Enhancing weed detection using UAV imagery and deep learning with weather-driven domain adaptation. *Computers and Electronics in Agriculture*, 237, 110673. <https://doi.org/10.1016/j.compag.2025.110673>

- [19] Yazdinejad, A., Zolfaghari, B., Azmoodeh, A., Dehghantanha, A., Karimipour, H., Fraser, E., ... & Duncan, E. (2021). A review on security of smart farming and precision agriculture: Security aspects, attacks, threats and countermeasures. *Applied Sciences*, 11(16), 7518. <https://doi.org/10.3390/app11167518>
- [20] Zhang, Y., Cai, H., Ye, J., Pan, F., Wu, S., Zhang, B., ... & Ma, R. (2025). Exploiting adversarial style for generalized and robust weed segmentation in rice paddy field. *Frontiers in Plant Science*, 16, 1703811. <https://doi.org/10.3389/fpls.2025.1703811>
- [21] Zhao, H., & Wang, Y. (2025). Deep learning-based approaches for weed detection in crops. *Frontiers in Plant Science*, 16, 1746406. <https://doi.org/10.3389/fpls.2025.1746406>

## Authors Biography



**Dr.J. Justina Michael** is an Assistant Professor in the Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai. With over 13+ years of academic and research experience, she has served as a Research Scholar, Assistant Professor, and Technical Trainer, contributing extensively to engineering education and professional training. She earned her Ph.D. in Computer Science and Engineering from SRM Institute of Science and Technology (2021–2025), and holds an M.E. in Computer Science and Engineering (2010) and a B.Tech. in Information Technology (2008) from Anna University. Dr. Justina is certified in High Impact Teaching Skills by WIPRO–ISTE and has mentored thousands of students. Her research focuses on Artificial Intelligence applications in agriculture, particularly deep learning architectures for weed identification aimed at enhancing sustainable farming practices. Her work has received notable recognition, including the ISWS Student Travel Grant Award (2022), Oral Speaker Distinction at the 3rd International Weed Conference (2022), Best Poster Presentation Award at DPRC 2022, and the Gold Medal for Research Excellence (2024) at SRMIST. Her research interests include Deep Learning, Artificial Intelligence, Computer Vision, and intelligent systems for societal impact.



**Dr.H. Mary Shyni** is an Assistant Professor at SRM Institute of Science and Technology, Vadapalani, Chennai, with over nine years of academic and research experience. She completed her Ph.D. at SRM Institute of Science and Technology (2021–2025), focusing on enhancing feature extraction for lung disease classification and segmentation from chest X-ray images using advanced deep learning architectures. She holds an M.Tech in Laser and Electro-Optical Engineering and a B.E. in Electronics and Communication Engineering from Anna University, Guindy. Dr. Shyni's research expertise lies in Machine Learning, Deep Learning, medical image analysis, and computer vision. She has published SCI- and Scopus-indexed journal articles, Springer book chapters, and IEEE conference papers, with significant citation impact. Her notable contributions include the development of PulmonNet and PulmonU-Net models for pulmonary disease detection and segmentation. She is also the inventor of an Indian patent on tuberculosis severity estimation using deep learning techniques. Recognized for research excellence, she received the Best Paper Award for her work on COVID-19 detection from chest X-rays. Her research interests include AI-driven healthcare solutions and intelligent diagnostic systems.



**G. Praveen Kumar** is an Assistant Professor, Research Scholar, and Technical Trainer with over 14 years of academic and professional experience in Computer Science and Engineering. He is currently serving as an Assistant Professor at St. Joseph's Institute of Technology, Chennai, and is pursuing his Ph.D. in Computer Science and Engineering at Anna University, Chennai. He completed his M.E. in Computer Science and Engineering from Anna University in 2010 and his B.Tech. in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad, in 2008. Mr. Praveen Kumar possesses strong expertise in programming languages such as Java, Python, C, and C++. He has also contributed to research in areas such as cloud computing, mobile cloud learning, iris recognition, and privacy-enhanced location-based services. Recognized for his teaching excellence, he has received the Best Faculty Award and Certificate of Proficiency for securing 100% results. His research interests include Cyber Security, Artificial Intelligence, Cloud Computing, and Intelligent Computing Systems.



**Dr.U. Sakthivelu** is an Assistant Professor in the Department of Computer Science and Engineering at Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India. He holds a Ph.D. in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai, along with Master's and Bachelor's Engineering degrees from Annamalai University. He also earned an MBA in Human Resource Management, reflecting his interdisciplinary academic background. His research interests include Machine Learning, Deep Learning, Cyber Security, and Intelligent Systems. Dr. Sakthivelu has published extensively in reputed international journals and conferences indexed in IEEE, Springer, and other leading platforms, contributing significantly to areas such as advanced persistent threat detection, cyber threat intelligence, healthcare applications, and optimization-driven deep learning models. He is also an innovator with patented contributions in AI-based security robotics and IoMT-enabled smart healthcare systems. With strong expertise in research, academic leadership, and collaborative projects, Dr. Sakthivelu actively engages in advancing intelligent computing solutions for real-world challenges.