Physical Education, Health and Social Sciences
https://journal-of-social-education.org
E-ISSN: 2958-5996
P-ISSN: 2958-5988

RGB Vegetation Index-Based Pixel-Level Classification of Crops and Weeds Using Machine Learning

Sidra Soomro[1], Zaid Hussain Dahar[2], Naveed Ahmed[3]

[1] Department of Software Engineering, Mehran University of Engineering & Technology Jamshoro, Sindh Pakistan. Email: sidrasoomro14@gmail.com
[2,3] Department of Computer Science, The Shaikh Ayaz University, Shikarpur, Sindh Pakistan. Email: zaiddahar1@gmail.com, ahmednaveed155@gmail.com

DOI: https://doi.org/10.63163/jpehss.v4i1.1291

Abstract
One of the primary causes of increasing poverty among farmers is the uncontrolled and excessive growth of weeds, which directly impacts crop yields. Over time, various surveys and techniques remained advance and implemented to differentiate weeds from crops in order to autonomously eliminate or manage them. Numerous approaches have been employed, including color-based, threshold-based, and machine learning-based methods. This study presents a high-quality, pixel-annotated agricultural dataset comprising of 60 RGB field images utilized for crop–weed–soil segmentation, containing over 75 million labeled pixels. Each image features manually created masks that distinguish soil, crops, and weeds. To address class imbalance, 70,000 samples per class were selected. The dataset is divide into training (60%), validation (20%), and test (20%) sets to enable robust evaluation of machine learning models for pixel-level crop and weed classification, aiming to address this critical issue that contributes to significant mental and physical stress for farmers, as well as substantial financial losses due to wasted cultivated land and inefficient use of labor. In particular, this review focuses on machine learning-based methods applied to evaluating their effectiveness in detecting and classifying weeds through advanced algorithms and model parameters.

Keywords: Identification of weeds, Machine Learning, Random Forest, Image Processing, SVM

Introduction

By 2050, there will likely be nine billion people on the planet, and agricultural food production would need to double to keep up with demand. [1], [2]. However, the primary causes that agriculture faces are weeds and a few other issues and difficulties. [3]. numerous research have been conducted to eliminate weeds that develop in fields and share resources with the main crop, causing the main crop to lose its intended yield [4], [5]. Dust moistness, nutrients, and sun radiation are used with crops in order to reduce weeds [6]. In order to reduce crop from weed, must take major step in the earlier stages to avoid huge loss of production. However, weeds negatively impact crops by causing them to share space, light, water, and nutrients; increasing production costs; making harvesting more difficult; degrading product quality; and devaluing the commercial cultivated areas. Unfortunately, farmers facing severe issue is weed since they threaten their capacity to produce high-quality food at a reasonable price. Weed competition and crop production loss are closely related. Because manual weeding is labor-intensive and challenging, it is not practical. Comparably effective, mechanical weeding is unable to eradicate intra-row weeds and occasionally, due to human error, may harm the primary crop. Because herbicides are so effective at reducing weeds, they are most frequently used [7]. Herbicide waste and environmental contamination result from the use of herbicides in all areas of agriculture [8]. The same effect as a full dosage can be obtained in low-density weed areas by using half of the amount [9]. To control weed the best method is to adopt is chemical weeding. The

two approaches that are regularly employed in Pakistan are manual spraying by farmers, which has an adverse effect on their health, and tractor-assisted spraying, which leads to high herbicide expenses and unintentional environmental contamination. Modern technology must be used to spray herbicides autonomously in order to prevent contamination and reduce herbicide loss. The initial stage in this process is to effectively and independently classify soil, crops, and weeds. There have been numerous attempts at image processing thus far. From Past analysis, there have been three primary tactics towards managing this distinction amid crops and weeds. While learning-based tactics offer greater precision, color and threshold-based tactics lose accuracy trendy extreme light conditions. However, SVM and ML are the two best methods to get accuracy. Machine learning has been the most popular approach since it offers the highest accuracy. Machine learning-based practices that have previously been used to weed and crop categorization or weed detection are covered in this review study.

Machine Learning

Machine learning (ML) is a branch of artificial intelligence in which system learn from data instead of fixed rules. Like an image-based agriculture tasks, ML is always used to separate different on pixel categories such as soil, crop, and weed. Instead of relying too deep multi-layer neural networks, traditional ML use methods depend mostly on best feature extraction. Features like RGB color values, vegetation indices, or more texture measures are calculated first, and the model then learns from how these features relate to each class. Models like Random Forest, SVM, and KNN work well when the dataset is not very large, which is common in agricultural research. These algorithms are generally easier to train, faster to test, and require less computing power than deep learning. One more advantage is that their decisions are simple to explain, since the model depends on clear input features rather than hidden layers. Sometimes, the performance of ML depends heavily on how the features are designed. It check the image quality if it is low, if plants overlap, or if the lighting varies a lot, the extracted features may not represent the crop or weed clearly. In such cases, the model might confuse visually similar pixels. Even with these limitations, ML models remains useful always for crop–weed classification, especially when the dataset is limited and a lightweight, interpretable approach is preferred.

Literature Review

This section primarily focuses on recent research associated to the classification of weeds in agriculture. Machine learning and digital image processing techniques play a vital role in this domain. In the present study, machine learning algorithms combined with image analysis methods are employed to effectively classify weeds using imagery captured by unmanned aerial vehicles [10-13]. Moreover, according to recent surveys, machine learning algorithms provide more accurate and efficient results for complex datasets compared to traditional methodologies [14-16]. Because of its broad performance and quick operation, the RF classifier is one of these machine learning algorithms that is quickly gaining popularity for remote sensing applications [14–16, 13]. RF has proven to be advantageous for agricultural and high resolution UAV picture categorization [3–5, 13]. A land cover map was created by Brinkhoff et al. [17] to identify and categorise perennial crops spanning more than 6200 km2 in the Riverina region of Australia's New South Wales. Their approach involved combining supervised SVM classification with object-based image examination approaches to improve precision. They discovered that the accuracy for an object count with twelve number of classes was 84.8% overall and 90.9% when weighted according to item area. According to these findings, precise maps of land cover across intensive perennial cropping areas might be created consuming a temporal series of medium resolution remote sensing measurements.

Alam et al. [3] developed a real-time computer vision-based system to differentiate between weeds and crops using an RF classifier. The classification model is trained on the authors' own dataset and

tested on ground data. Additionally, they formed a fluid flow control system founded on pulse width modulation, which customs information from the vision-based system to operate a piece of equipment that sprays a specific quantity of agrochemical. Consequently, the authors' real-time vision-based pesticide spraying system's efficacy was shown. An SVM approach was presented by Faisal et al. [18] to recognise weeds in images of chilli fields. Finding out how well the classifier using the SVM performed in a holistic weed management approach was the aim of their study. Photographs were taken of five distinct weeds from Bangladeshi chilli farms. Then, in order to extract attributes and distinguish the plants from the ground, these images were separated through an overall thresholding-based binarization technique. The fourteen characteristics of each image were separated into three groups: moment invariants, colour features, and shape features. Finally, an SVM classifier is used to identify weeds. According to their test results, the SVM's overall accuracy on 224 test photos was 97%. To distinguish the weeds from the remarkably identical sugar beets, the researchers of [19] used a range of form characteristics to create patterns that served as a weed identification method for sugar beet farming. The study's photos were shot on the sugar beetroot fields at Shiraz University. These pictures were processed using the MATLAB toolbox. The researchers investigated several shape features, including shape factors, moment invariants, and Fourier descriptors, in order to distinguish between sugar beets and weeds. Then KNN and SVM classifiers were applied, with 92.92% and 95% overall accuracies, respectively.

In [20], a weed detection method was presented that creates a monochromatic image and uses a colour index-based histogram to classify soil, soybean, and weed (broadleaf) classes using colour indices as a feature. Images were scaled to an interval between 0 and 255 to create greyscale images. Image histograms were then created and normalized, then BPNN and SVM classifiers were trained using these histograms. Finding an alternative feature vector that combines simple computation processes with a high weed detection rate was the aim of this work. Overall accuracies for SVM and BPNN using this approach were 95.078% and 96.601%, respectively.

Methodology

In this study, we employed a high-quality, pixel-annotated agricultural dataset for crop–weed– soil segmentation. The dataset consists of 60 RGB field images captured under natural illumination using a ground-based camera system. Each image is accompanied by manually created pixel-level annotation masks distinguishing soil, crops, and weeds, resulting in over 75 million labeled pixels.

Dataset Extraction

In this section verifies dataset extraction and correct folder structure, ensuring all images, masks, and annotations are loaded properly. It is placed in the Dataset Description section and establishes the foundation for reproducibility and preprocessing integrity.

```
print("Using dataset:", dataset_zip_path)


!rm -rf /content/dataset-master/

!unzip -q "$dataset_zip_path" -d /content/

print("Dataset extracted successfully!")
```

```
import os

root = "/content/dataset-master"

print("Folders:", os.listdir(root))
print("Images:", len(os.listdir(root + '/images')))
print("Masks:", len(os.listdir(root + '/masks')))
print("Annotations:", len(os.listdir(root + '/annotations')))
```

```
Folders: ['images', 'annotations', 'README.md', 'train_test_split.yaml', 'masks']
Images: 60
Masks: 60
Annotations: 120
```

Fig. 1: Import Data Set

Original Pixel-Level Class Distribution

The plot displays the original pixel-class imbalance, highlighting soil dominance over crop and weed pixels. This imbalance justifies applying balanced sampling during preprocessing.

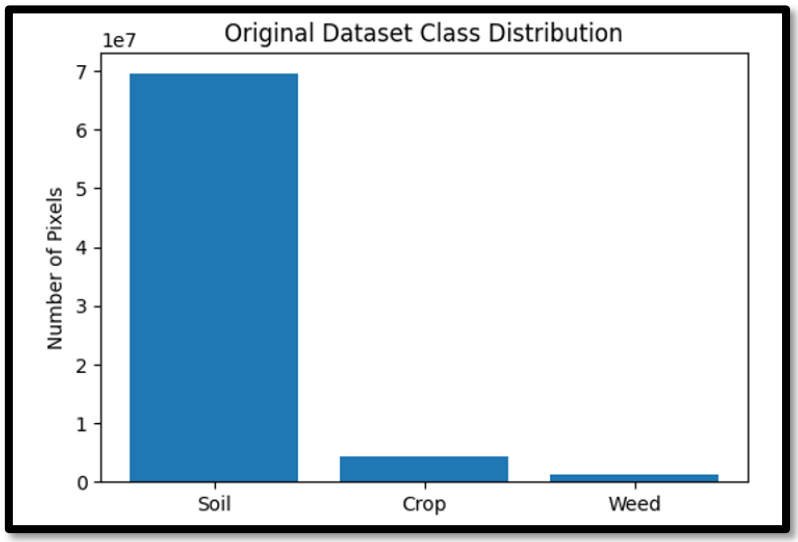

Fig. 2: Original Dataset Class Distribution

Data Preprocessing & Sample Rgb Image

It shows a sample RGB field image and its ground truth mask. It demonstrates pixel-level labeling for soil, crop, and weed regions, essential for understanding the complexity of the classification task. The dataset exhibits significant class imbalance, with soil pixels overwhelmingly dominating the dataset (~73.9 million pixels) compared to crops and weeds (~1.21 million pixels for weeds). To address this issue, balanced pixel resampling performed, selecting 70,000 samples per class (soil,

crop, and weed). This step ensured that each class contributed equally during model training and prevented bias toward the majority class.

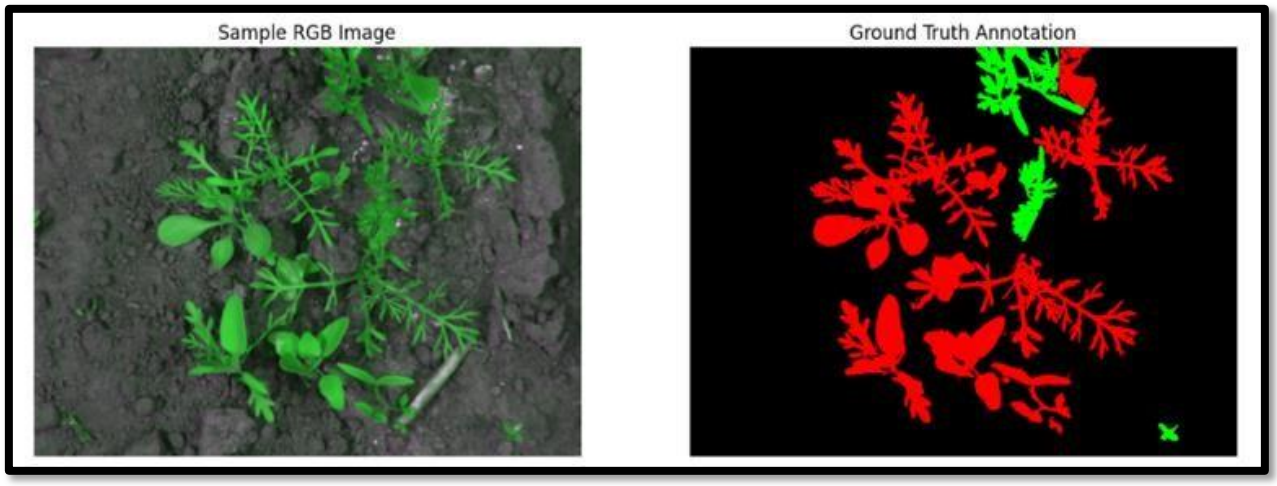

Fig. 3: Sample Input RGB & Corresponding ground truth annotation

Train–Validation–Test

The balanced dataset was divided into three subsets to enable robust evaluation of machine learning classifiers:

- **Training set:** 60% of the balanced samples (126,000 pixels)
- **Validation set:** 20% of the balanced samples (42,000 pixels)
- **Test set:** 20% of the balanced samples (42,000 pixels)

This split preserves equal class representation across all subsets, facilitating fair and consistent model evaluation.

```
from sklearn.model_selection import train_test_split

# 60% train, 20% val, 20% test
X_train, X_temp, y_train, y_temp = train_test_split(
    X_small, y_small,
    test_size=0.4,
    random_state=42,
    stratify=y_small
)

X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp,
    test_size=0.5,
    random_state=42,
    stratify=y_temp
)

print("Train:", X_train.shape, y_train.shape)
print("Validation:", X_val.shape, y_val.shape)
print("Test:", X_test.shape, y_test.shape)
```

```
Train: (126000, 7) (126000,)
Validation: (42000, 7) (42000,)
Test: (42000, 7) (42000,)
```

Fig. 4: SK Learn Model Selection

Feature Extraction

To enhance discrimination between crops, weeds, and soil, RGB-based vegetation indices were computed for each pixel. These indices capture color and reflectance characteristics associated with vegetation, enabling machine learning models to differentiate crop plants from weeds effectively.

Feature vectors for each pixel were constructed using these indices and normalized prior to model training.

Machine Learning Models

Classical machine learning classifiers including Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN) trained on the extracted pixel-level features. Hyper parameters for each model optimized using the validation set, ensuring optimal performance while avoiding over fitting. Each algorithm was trained to capture different aspects of the feature space: RF leveraged ensemble-based decision trees to handle non-linear patterns, SVM aimed to find an optimal separating hyper plane for high-dimensional feature distributions, and KNN classified each pixel based on similarity to its nearest neighbors. To ensure fair comparison and robust performance, model hyper parameters were systematically optimized using the validation set. This involved tuning parameters such as the number of trees for RF, kernel type and regularization for SVM, and the value of k for KNN. The optimization process minimized the risk of over fitting while maximizing generalization capability across unseen samples. These classical models provided useful insights into the discriminative power of the hand-crafted features and served as benchmarks for comparison with deep learning-based approaches.

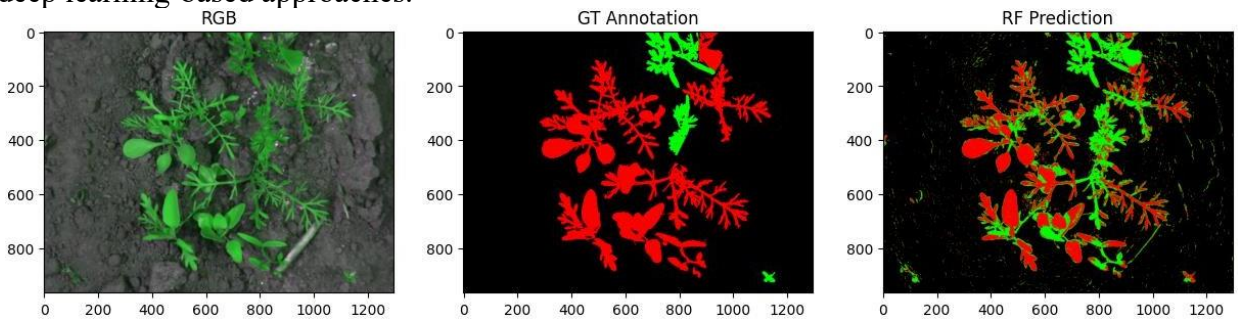

Fig. 5: RGB, GT Annotation & RF Prediction

Evaluation Metrics

Model performance was evaluated on the test set using standard metrics for multi-class classification, including overall accuracy, precision, recall, and F1-score for each class. These metrics provide a comprehensive assessment of the classifiers' ability to correctly identify crops, weeds, and soil at the pixel level. Precision and recall helped quantify the degree to which models avoided false positives and false negatives for each class, whereas the F1-score provided a balanced measure of both. In addition, overall accuracy captured the general classification effectiveness across the entire dataset. This multi-metric evaluation framework ensured that the models is not only accurate but also robust across classes with varying sample distributions. The detailed performance analysis further enabled meaningful comparisons between classical machine learning algorithms and more advanced deep learning approaches.

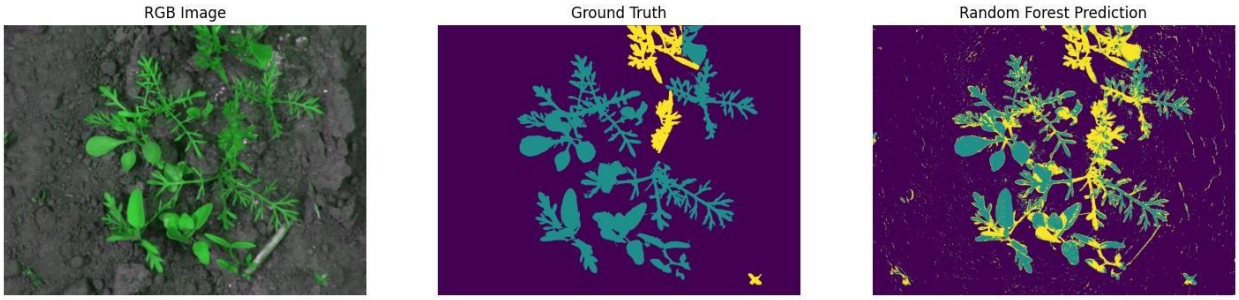

Fig. 6: RGB Image, Ground Truth & RF Prediction

Per-Class Performance Analysis

The per-class F1-score and IoU results highlight consistent trends across the three machine learning models. All models achieve their highest performance on the Soil class, with F1 and IoU values close to 1.0, indicating that soil pixels are the easiest to classify due to their distinct visual features. Performance decreases for the Crop and Weed classes, where class boundaries are more complex. Among the models, Random Forest (RF) consistently achieves the highest F1 and IoU scores for all classes, demonstrating its stronger generalization capability. SVM and KNN show moderate performance, with notably lower scores for the Crop and Weed classes, reflecting their limitations in handling intra-class variability. Overall, the results confirm that RF is the most effective classical model for distinguishing crop, weed, and soil pixels in the dataset.

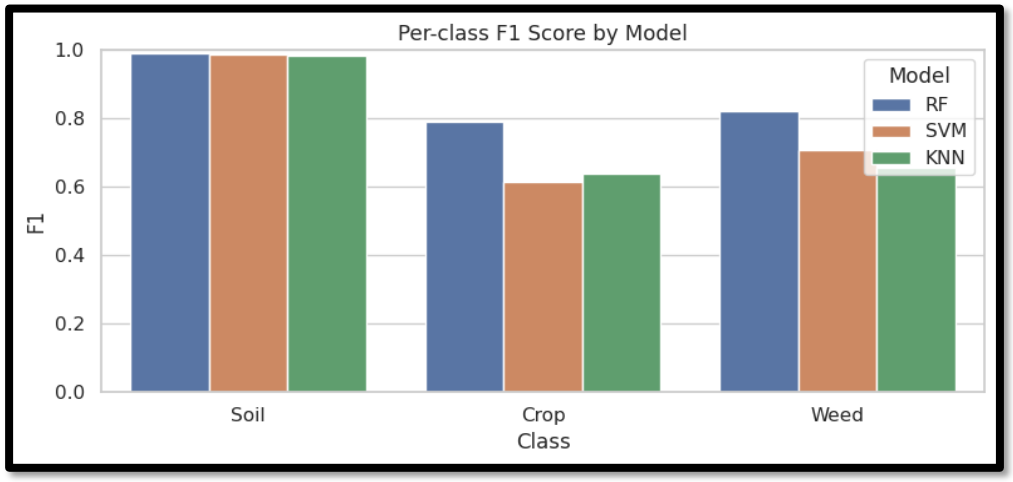

Fig. 7: Per-class F1 Score by Model

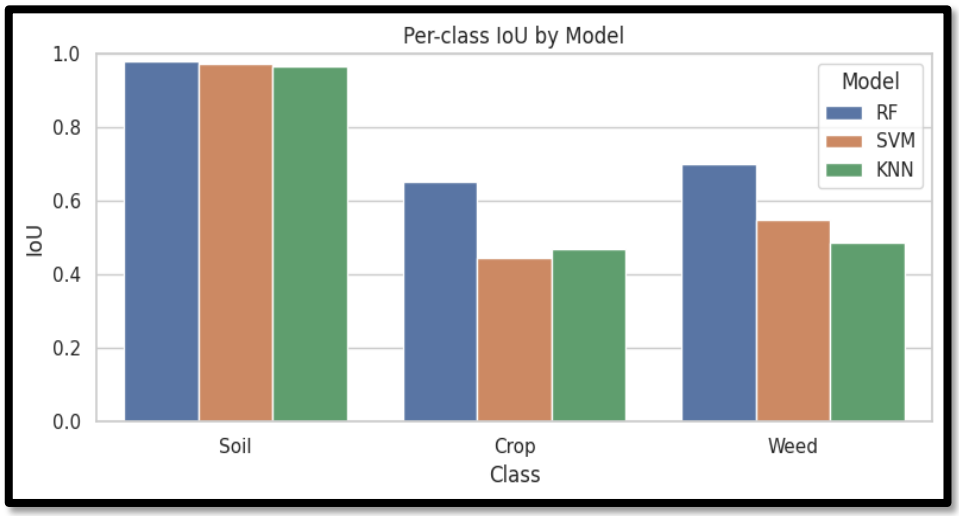

Fig. 8: Per-class IoU by Model

Random Forest Confusion Matrix

Confusion matrix for the Random Forest classifier showing strong soil detection and moderate weed–crop separation. It supports RF's position as the best-performing ML model

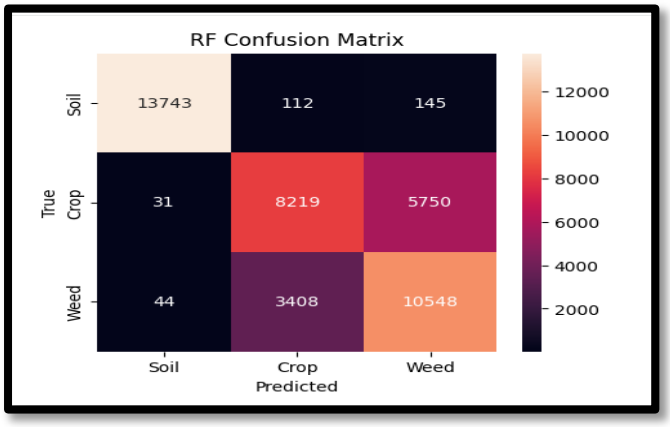

Fig. 9: RF Confusion Matrix

SVM Confusion Matrix

SVM confusion matrix indicating good soil accuracy but weak weed detection. Helps analyze SVM's performance behavior across classes.

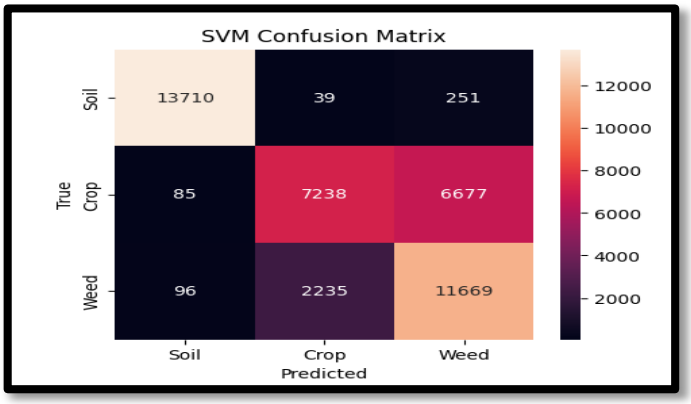

Fig. 9: SVM Confusion Matrix

KNN Confusion Matrix

KNN confusion matrix revealing noise sensitivity and moderate multiclass performance, assisting in comparative evaluation across traditional ML models.

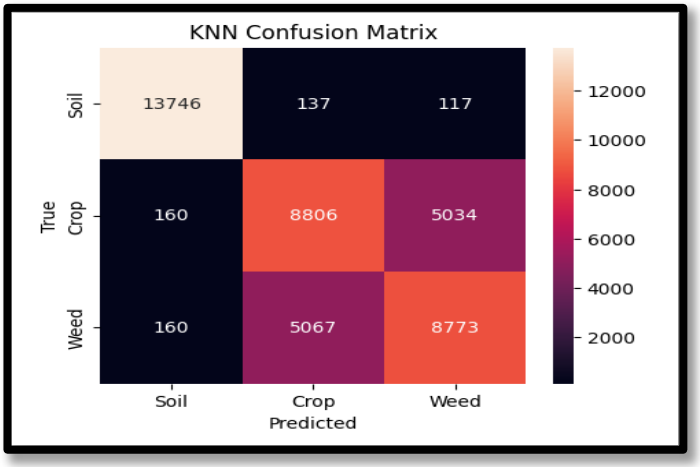

Fig. 10: KNN Confusion Matrix

Results

The following summarizing Accuracy, Precision, Recall, and F1-score for RF, SVM, and KNN. This supports quantitative comparison of classifiers.


```
{'RF': {'Accuracy': 0.7864761904761904,
  'Precision (Macro)': 0.7919551700551156,
  'Recall (Macro)': 0.7864761904761904,
  'F1 (Macro)': 0.7854636001564349},
 'SVM': {'Accuracy': 0.7765952380952381,
  'Precision (Macro)': 0.791790111298044,
  'Recall (Macro)': 0.7765952380952381,
  'F1 (Macro)': 0.7715846318231248},
 'KNN': {'Accuracy': 0.7458333333333333,
  'Precision (Macro)': 0.7452881139483206,
  'Recall (Macro)': 0.7458333333333332,
  'F1 (Macro)': 0.7455573397660484}}
```


Fig. 11: Accuracy, Precision, Recall, and F1-score for RF, SVM, and KNN.

Results of Crop Vs Weed Binary Evaluation

Binary evaluation focusing on crop and weed classification (soil removed). Valuable for real-world applications like precision spraying.

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)
RF	0.7865	0.792	0.7865	0.7855
SVM	0.7766	0.7918	0.7766	0.7716
KNN	0.7458	0.7453	0.7458	0.7456

Table. 1: Accuracy, Precision, Recall, and F1 for Model RF, SVM, and KNN.

Final Visual Prediction Comparison

Final qualitative comparison showing predictions from RF, SVM, and KNN alongside ground truth. This figure visually summarizes model strengths and weaknesses.

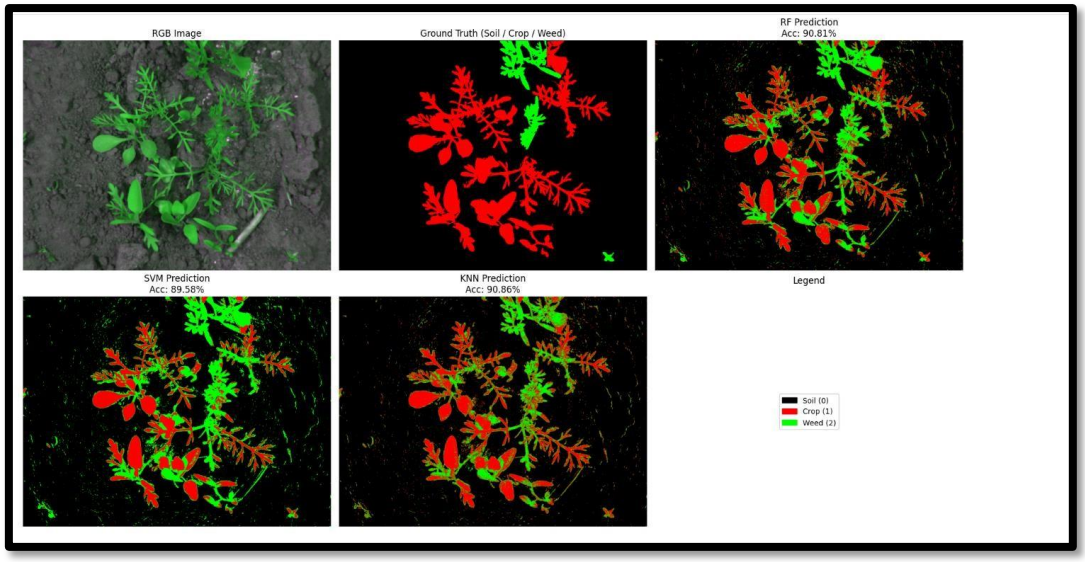

Fig. 12: RF, SVM, and KNN Prediction With Accuracy.

As discussed earlier, Random Forest (RF), Support Vector Machine (SVM), and KNN have been extensively adopted for the classification of crops and weeds. Therefore, in this study, we selected these three algorithms to evaluate their performance and compare their classification results. The study area categorized into three distinct classes: crops, weeds, and soil.

Comparative Analysis of RF, SVM, and KNN

The figure presents a comparative analysis of three classical machine learning models Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN) based on their global (macro-averaged) performance metrics. The bar chart evaluates four standard multi-class classification metrics: Accuracy, Precision (Macro), Recall (Macro), and F1-score (Macro). Each metric is displayed along the x-axis, while the y-axis represents the corresponding performance values ranging from 0.0 to 1.0. Across all metrics, the Random Forest model consistently achieves the highest scores, with values slightly above 0.85 for accuracy, precision, recall, and F1-score, indicating strong and balanced classification performance. The SVM model performs moderately well, achieving scores around 0.77–0.78 across all metrics, demonstrating reliable but comparatively lower effectiveness than RF. The KNN model shows slightly lower results than SVM, with all metric values around 0.75, suggesting that while it provides reasonable performance, it is less robust, likely due to sensitivity to feature scaling and local variations in pixel-level data. Overall, the visual comparison highlights that RF outperforms the other two algorithms across all evaluation metrics, demonstrating its superior ability to generalize from the extracted pixel-level features. The SVM and KNN models show comparable but lower performance, reflecting differences in their capacity to capture complex class boundaries in the crop–weed–soil segmentation task.

	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)	mIoU
RF	0.868500	0.874874	0.868500	0.868153	0.778207
SVM	0.774429	0.787022	0.774429	0.770393	0.655717
KNN	0.758952	0.760169	0.758952	0.759288	0.641467

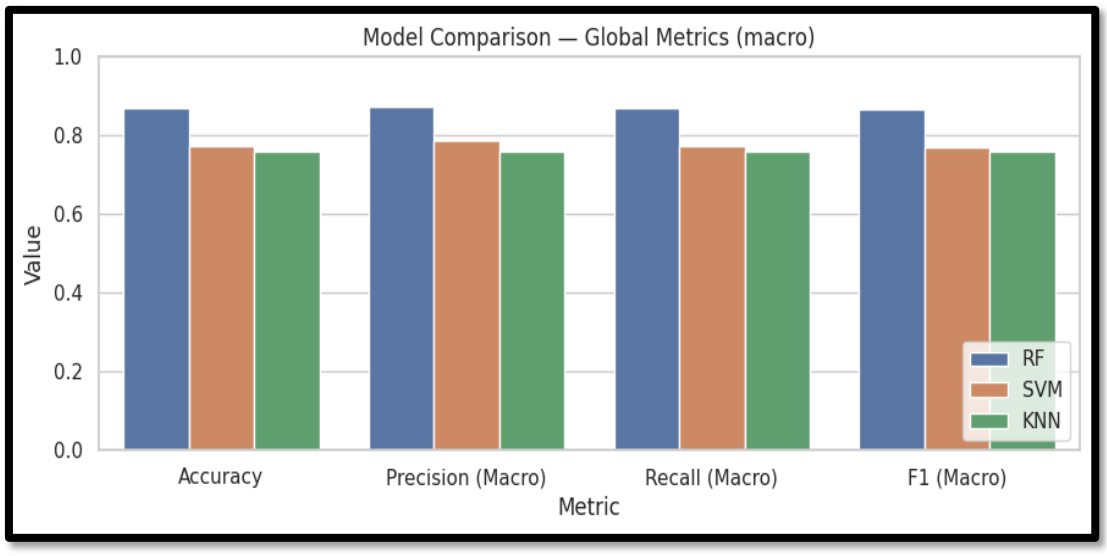

Fig. 13: Model Comparison.

ROC Curves for One-vs-Rest Multi-Class Classification Using Random Forest, SVM, and KNN Models

The figure shows Receiver Operating Characteristic (ROC) curves for a multi-class classification problem (Soil, Crop, Weed) using three models Random Forest (RF), Support Vector Machine

(SVM), and k-Nearest Neighbors (KNN). Each class is evaluated using the One-vs-Rest strategy, where the selected class is treated as "positive" and all others as "negative." The x-axis (FPR) whispers how often the model raises false alarms. The y-axis (TPR) shows how often the model catches the true positives. A curve traveling toward the top-left corner signals stronger performance, like a compass pointing to reliable decision-making. The diagonal line represents random guessing.

Random Forest

RF-Soil (AUC = 1.00): A flawless arc, hugging the top boundary. The model separates Soil from other classes with complete confidence. RF-Crop (AUC = 0.94) and RF-Weed (AUC = 0.95): Both show excellent discrimination.

SVM

SVM-Soil (AUC = 0.99): Nearly perfect. SVM-Crop (AUC = 0.87) and SVM-Weed (AUC = 0.86): Good performance, but the curves wander slightly below RF, indicating more overlap in features.

KNN

KNN-Soil (AUC = 0.99): Surprisingly strong. KNN-Crop (AUC = 0.83) and KNN-Weed (AUC = 0.84): Adequate but clearly the least assertive among the models.

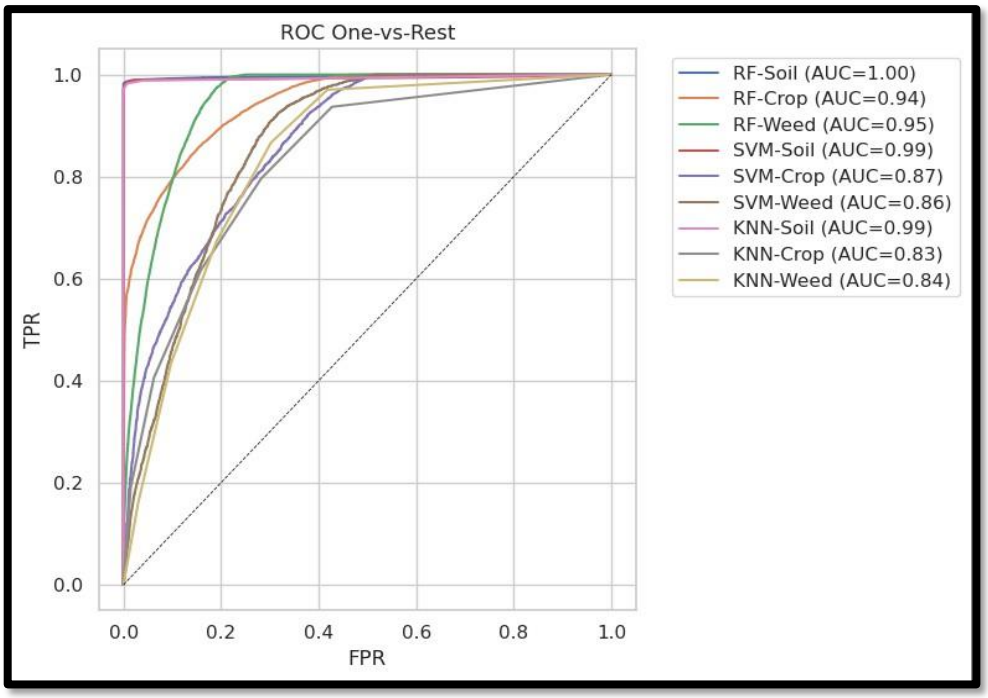

Fig. 14: ROC One-vs-Rest

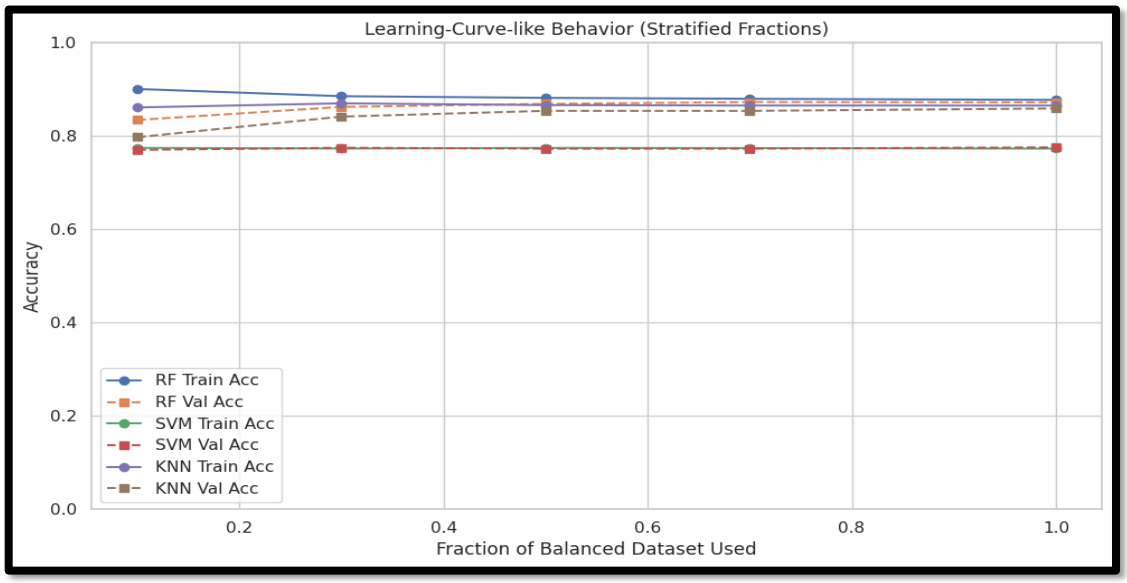

Fig. 15: Fraction of Balanced Dataset

Conclusion

Machine learning represents an innovative advancement in autonomous weeding, offering greater accuracy compared to traditional methods. It addresses a critical research gap by enabling the automatic identification of weeds and targeted treatment across various crop types. With the automated application of herbicides, farmers can achieve improved crop yields and enhanced precision in weed management. Additionally, the selective use of herbicides helps minimize soil contamination, promoting more sustainable agricultural practices.

References

[1] Cheng, B., & Matson, E. T. (2015). A feature-based machine learning agent for automatic rice and weed discrimination. In *Proceedings of the 13th International Conference on Intelligent Autonomous Systems* (pp. 517–527). Springer.

[2] Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science, 21*(2), 110–124.

[3] Lee, W. S., Alchanatis, V., Yang, C., Hirafuji, M., Moshou, D., & Li, C. (2010). Sensing technologies for precision specialty crop production. *Computers and Electronics in Agriculture, 74*, 2–33.

[4] Berge, T. W., Aastveit, A. H., & Fykse, H. (2008). Evaluation of an algorithm for automatic detection of broad-leaved weeds in spring cereals. *Precision Agriculture, 9*, 391–405.

[5] Hamuda, E., Glavin, M., & Jones, E. (2016). A survey of image processing techniques for plant extraction and segmentation in the field. *Computers and Electronics in Agriculture, 125*, 184–199.

[6] Bakker, H. (2012). *Sugar cane cultivation and management*. Elsevier.

[7] Ali, A., Streibig, J. C., Christensen, S., & Anderson, C. (2014). Image-based thresholds for weeds in maize fields. *Weed Research, 55*(1), 26–33.

[8] Qi, P., Luo, X. H., & Zhang, D. S. (2009). Weed recognition based on digital image processing in wheat field. *Journal of Xinhua University (Natural Science Edition)*, 136–137.

[9] Wu, L. L., Liu, J. Y., Wen, Y. X., & Deng, X. Y. (2009). Weed identification method based on SVM in the cornfield. *Transactions of the Chinese Society of Agricultural Machinery, 40*(1), 162–166.

[10] Tian, H., Wang, T., Liu, Y., Qiao, X., & Li, Y. (2020). Computer vision technology in

agricultural automation—A review. *Information Processing in Agriculture, 7*, 1–19.

[11] Wang, A., Zhang, W., & Wei, X. (2019). A review on weed detection using ground-based machine vision and image processing techniques. *Computers and Electronics in Agriculture, 158*, 226–240.

[12] Herrmann, I., Shapira, U., Kinast, S., Karnieli, A., & Bonfil, D. (2013). Ground-level hyperspectral imagery for detecting weeds in wheat fields. *Precision Agriculture, 14*, 637–659.

[13] Weis, M., Gutjahr, C., Ayala, V. R., Gerhards, R., Ritter, C., & Schölderle, F. (2008). Precision farming for weed management: Techniques. *Gesunde Pflanzen, 60*, 171–181.

[14] Alam, M., Alam, M. S., Roman, M., Tufail, M., Khan, M. U., & Khan, M. T. (2020). Real- time machine-learning-based crop/weed detection and classification for variable-rate spraying in precision agriculture. In *Proceedings of the 2020 7th International Conference on Electrical and Electronics Engineering (ICEEE)* (pp. 273–280). Antalya, Turkey.

[15] Gao, J., Nuyttens, D., Lootens, P., He, Y., & Pieters, J. G. (2018). Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery. *Biosystems Engineering, 170*, 39–50.

[16] De Castro, A. I., Torres-Sánchez, J., Peña, J. M., Jiménez-Brenes, F. M., Csillik, O., & López-Granados, F. (2018). An automatic random forest-OBIA algorithm for early weed mapping between and within crop rows using UAV imagery. *Remote Sensing, 10*, 285.

[17] Brinkhoff, J., Vardanega, J., & Robson, A. J. (2020). Land cover classification of nine perennial crops using Sentinel-1 and -2 data. *Remote Sensing, 12*, 96.

[18] Ahmed, F., Al-Mamun, H. A., Bari, A. H., Hossain, E., & Kwan, P. (2012). Classification of crops and weeds from digital images: A support vector machine approach. *Crop Protection, 40*, 98–104.

[19] Bakhshipour, A., & Jafari, A. (2018). Evaluation of support vector machine and artificial neural networks in weed detection using shape features. *Computers and Electronics in Agriculture, 145*, 153–160.

[20] Abouzahir, S., Sadik, M., & Sabir, E. (2018). Enhanced approach for weeds species detection using machine vision. In *Proceedings of the 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)* (pp. 1–6). Kenitra, Morocco.