Evaluating deep learning models for effective weed classification in agricultural images

Bernardo Silva Ribeiro Duarte[1], Elisa Ribeiro Gonçalves[2], Pedro Henrique Campos Moreira[3]
[1]Instituto de Ciências Exatas e Técnologicas - Universidade Federal de Viçosa - Campus Rio Paranaíba (UFV-CRP)
Rodovia MG230, Km7, Caixa Postal 22 - 38.810-000 - Rio Paranaíba - MG - Brasil
bernardo.duarte@ufv.br, elisa.r.goncalves@ufv.br, pedro.henrique.moreira@ufv.br

Abstract—**Effective weed management is crucial for maximizing agricultural productivity and minimizing crop losses. Traditional methods for weed detection and classification often suffer from errors and delays, particularly in the absence of specialists. Deep learning technologies offer a promising alternative for automating these tasks, potentially enhancing both accuracy and efficiency. This study compares three advanced deep learning architectures, ResNet-50, EfficientNet V2, and Vision Transformers (ViT), for classifying weed species using the DeepWeeds dataset. We explore the effects of data augmentation on model performance, evaluating each model based on accuracy, precision, recall, and F1 score. Our results demonstrate that data augmentation significantly improves model performance. EfficientNet V2 achieved the highest performance across all metrics, with a peak accuracy of 0.9703. This research provides valuable insights into selecting effective architectures and training strategies for more accurate weed detection in agricultural applications.**

Index Terms—**Deep Learning, Weed Classification, Data Augmentation, CNNs.**

I. Introduction

Agriculture is a fundamental productive sector that will remain crucial to human activity for centuries to come. Over the past few decades, Brazil has advanced its agricultural sector to achieve global competitiveness, becoming a key player in the industry [1]. The country's success in agriculture is attributed to the adoption of advanced technologies, including new crop varieties adapted to diverse environmental conditions and precision agriculture techniques aimed at boosting productivity [2]. In the first quarter of 2023, Brazil's agricultural sector experienced a remarkable growth of 21.6%, largely driven by a record soybean harvest, which contributed significantly to a 1.9% increase in the nation's GDP [3]. Despite these achievements, agricultural productivity faces significant challenges, particularly from competition with weeds.

Weeds are one of the main biotic limitations to agricultural production, causing significant losses of up to 31.5% in plant yield [4]. They compete with crops for light, water, nutrients, and space, directly harming agricultural productivity. Additionally, they can harbor insects and pathogens that attack crops, and their presence leads to the degradation of native habitats, threatening local plants and animals [5]. Traditional control methods, such as intensive herbicide use, besides being costly, negatively impact the ecosystem and human health [6].

Given the rapid advancement of agricultural technology, there is a growing interest in developing automated systems for identifying and classifying plant species, especially weeds, due to their impact on productivity and the environment. Computer vision and machine learning stand out as promising alternatives to improve the precision and speed of detection.

Deep learning techniques have shown great potential in automating the weed classification process, providing high accuracy in recognizing complex patterns [7]. These techniques are particularly effective in large-scale cultivation areas where manual control is unfeasible. However, their effectiveness depends on factors such as neural network architecture and data augmentation techniques, including variations in light and image angles, which are essential for increasing model robustness and ensuring applicability in diverse scenarios [8].

In this study, we compare three deep learning architectures (ResNet-50, EfficientNet V2, and ViT-b16) to classify weeds in agricultural images. We also evaluated the impact of training the models with and without data augmentation procedures. We trained and testes our models using the DeepWeeds dataset a widely used dataset for this type of classification task.

The paper is structured as follows: In Section II, we review related works in plant species classification. In Section III, we detail the dataset, deep learning architectures, and the experiment design. We present and discuss the results in Section IV. Finally, in Section V, we conclude the paper and suggest directions for future research.

II. Related works

Recent advances in deep learning and computer vision have driven the development of automatic systems for classifying weeds. Several studies have explored convolutional neural networks (CNNs) to identify plant species, including crops and weeds [9].

Olsen et al. [10] introduced DeepWeeds, a dataset containing 17,509 images of eight weed species found in northern Australia and images without weeds. Designed to support automated detection, this dataset was evaluated using Inception-v3 and ResNet-50 models, achieving up to 95% accuracy. The results demonstrate the dataset's potential to facilitate the development of robust deep learning models for automated weed management in precision agriculture.

Saleem et al. [11] evaluated single-stage convolutional neural network (CNN) models such as SSD and YOLO, as well as dual-stage networks such as Faster R-CNN, using the

DeepWeeds dataset. The faster R-CNN model with ResNet-101 achieved an average accuracy of 93.44%. Key improvements were achieved by employing advanced image-resizing techniques. Further enhancements were made through weight initialization methods and batch normalization, contributing an additional 1.82% improvement. The optimal selection of hyperparameters for the RMSProp optimizer resulted in a final boost in accuracy, demonstrating the significance of tuning these parameters for effective weed identification in complex agricultural environments.

Yang et al. [12] introduced a dissimilarity-based active learning (DBAL) method for embedded weed identification, effectively reducing the need for large labeled datasets by selecting representative samples. Their approach achieved 90.75% accuracy using only 32% of the DeepWeeds dataset and 98.97% accuracy with 27.8% of the Grass-Broadleaf dataset. Additionally, they compressed the model from 117.9 MB to 8.6 MB (a 92.7% compression ratio), with a minimal accuracy drop of 1%, and successfully deployed it on an NVIDIA Jetson AGX Xavier, achieving 192 fps. This demonstrates the potential of active learning and model compression in enabling real-time weed detection in resource-constrained environments.

Wang et al. [13] proposed a fine-grained weed recognition method based on the Swin Transformer architecture. This approach leverages two-stage transfer learning to address challenges in recognizing visually similar crops and weeds, achieving state-of-the-art performance. On the MWFI dataset, the model obtained a remarkable accuracy of 99.18%, outperforming traditional CNN models like VGG-16 and DenseNet-121. Additionally, the model demonstrated its effectiveness on the DeepWeeds dataset, proving its applicability in real-world agricultural environments.

Zhang et al. [14] proposed a multi-class weed recognition model combining CNNs and SVMs, tested on the DeepWeeds dataset. Their approach, using a ResNet-50-SVM architecture, achieved 97.6% accuracy, outperforming standalone CNN models such as ResNet-50 (96.1%), GoogLeNet (93.6%), and Densenet-121 (94.3%). The hybrid CNN-SVM method demonstrated superior performance by leveraging the feature extraction capabilities of CNNs alongside the classification precision of SVMs. This model proved effective in recognizing weeds in complex agricultural environments.

Zhang et al. [15] proposed a patch-based deep learning method for crop and weed recognition, splitting images into patches to train neural networks. Tested on datasets like DeepWeeds, Cotton weed, and Corn weed, the approach improved performance, especially with class imbalances. DenseNet201 achieved an F1 score of 98.49% on DeepWeeds and 100% on Cotton Tomato weed. The method also enhanced accuracy for minority classes in the Cotton weed dataset, addressing challenges of intra-class variability and inter-class similarity.

Belissent et al. [16] presented the transfer learning and zero-shot learning (ZSL) for the scalable detection and classification of weeds in images captured by unmanned aerial vehicles (UAVs) to optimize weed management and reduce herbicide use. Using the TomatoWeeds dataset, which contains images of three weed species, the authors evaluated the effectiveness of pre-trained residual networks on large datasets such as DeepWeeds. The pre-trained and refined ResNet50 model achieved an accuracy of 77.8%, demonstrating the potential of transfer learning in this domain.

In contrast to all previous works, our approach incorporates advanced data augmentation techniques and training based on fine-tuning. We also implement a learning rate scheduler (ReduceLROnPlateau) and early stopping to prevent overfitting.

III. Material and methods

A. Dataset

The DeepWeeds dataset [10][1] used in this study is a widely established resource for weed species classification. It consists of a total of 17,509 images, separated into nine distinct classes representing different weed species: *Chinee apple* (1,125 images), *Lantana* (1,064 images), *Negative* (9,106 images), *Parkinsonia* (1,031 images), *Parthenium* (1,022 images), *Prickly acacia* (1,062 images), *Rubber vine* (1,009 images), *Siam weed* (1,074 images), and *Snake weed* (1,016 images).

Each image in the dataset is labeled with the corresponding species, providing a rich data source for training deep learning models. Figure 1 shows an example of each class contained in this dataset.

B. Architectures

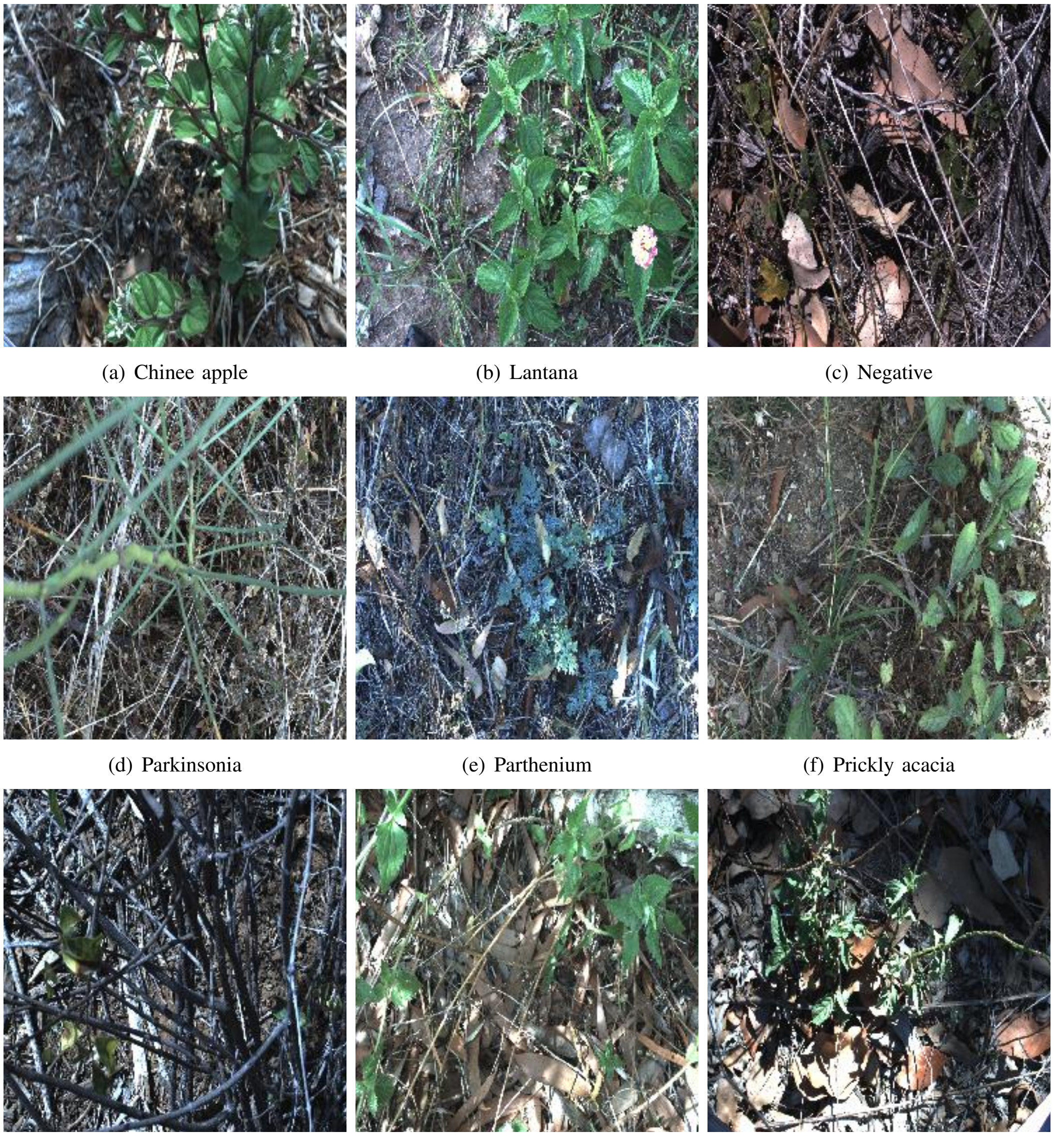
We selected three state-of-the-art deep learning architectures for this study: ResNet-50, EfficientNet V2, and ViT b 16.

ResNet-50 (Residual Network) [17] is renowned for introducing residual blocks, which employ shortcut connections to address the performance degradation issue in very deep networks. These connections allow gradients to flow directly between layers, mitigating the problem of vanishing gradients and facilitating the training of deep networks. As a result, ResNet-50 has proven to be highly effective in large-scale computer vision tasks, delivering strong performance in various benchmarks.

EfficientNetV2 [18], on the other hand, was designed to enhance both the accuracy and efficiency of convolutional networks. This architecture integrates Neural Architecture Search (NAS) with progressive learning, adjusting image size and regularization throughout the training process. Additionally, Fused-MBConv layers are utilized to accelerate training in the early layers. Building on the success of EfficientNet, EfficientNetV2 efficiently scales network dimensions (depth, width, and resolution) while maintaining fewer parameters, achieving remarkable results in computer vision.

Finally, Vision Transformer (ViT-b16) [19] departs from traditional CNNs, applying transformer architecture to image classification. Instead of using convolutions, ViT divides images into 16x16 pixel patches and processes them sequentially, similar to how transformers operate in natural language processing tasks. This approach allows ViT to capture long-range

[1] https://github.com/AlexOlsen/DeepWeeds



(a) Chinee apple (b) Lantana (c) Negative

(d) Parkinsonia (e) Parthenium (f) Prickly acacia

(g) Rubber vine (h) Siam weed (i) Snake weed

Fig. 1. One sample from each class of the Weeds dataset.

dependencies in images, a common limitation in CNNs. ViT-b16 has demonstrated excellent performance across various computer vision benchmarks, requiring significantly fewer computational resources for training on large datasets.

C. Data augmentation

Data augmentation techniques were employed to artificially increase the size and variability of the training data, thereby improving the model's generalization [20]. The augmentation strategies included random rotations, horizontal and vertical flips, zooming, and brightness adjustments. These transformations were applied during training to simulate various real-world conditions that the model might encounter, such as different lighting and viewing angles.

D. Experimental setup

First, we randomly split the Deep Weeds dataset into training, validation, and testing sets, with a stratified distribution of 70% for training, 15% for validation, and 15% for testing.

We fine-tuned pre-trained models provided with the torchvision library, using Adam optimizer, given its efficiency in handling sparse gradients and achieving faster convergence [21]. A batch size of 16 and an initial learning rate of 0.0001 was used for all experiments.

To mitigate overfitting, we employed the reduce the learning rate on plateaus strategy, which reduces the learning rate when the validation accuracy ceases to improve. Additionally, early stopping was utilized to terminate training if no enhancement in validation accuracy was observed over 10 consecutive epochs. This strategy helps avoid excessive training and overfitting. We capped the maximum number of training epochs at 200 to ensure sufficient time for model convergence while balancing computational efficiency.

E. Model evaluation

The model's performance was evaluated using the following metrics: accuracy, precision, recall, and F1-Score. The equations take into account the rates of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Accuracy measures the proportion of correct predictions (TP + TN) divided by the total number of instances evaluated, as described in Equation 1. It considers both the correctness of positive and negative cases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision refers to the proportion of correct positive predictions (TP) related to the total number of positive predictions (TP + FP), in accordance with Equation 2. This metric is useful in scenarios where we aim to minimize false positives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall measures the model's ability to identify all positive instances, i.e., correctly, the ratio between true positives (TP) and the total number of positive instances (TP + FN), as described in Equation 3. It is useful when we want to minimize false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score is the harmonic mean between precision and recall, providing a balanced metric that considers both measures (Equation 4). It is particularly useful when there is a need to balance precision and recall, especially when these metrics are equally relevant.

$$F1 - Score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

F. Computacional resourses

The experiments were conducted on a PC with a 3.00 GHz Core I5 CPU and 16 GB of RAM equipped with a GPU NVIDIA GEFORCE GTX 1050 ti with 4 GB of memory. We used Python 3.10 and PyTorch 2.2.2 [22] with CUDA 12.1 and torchvision 0.17.2. We also employed Scikit-learn 1.4.2 and Matplotlib 3.8.4. The environment was configured to optimize computational efficiency, allowing rapid training and evaluation of the models.

IV. Results and discussion

The results of our experiments are summarized in Table I, which compares the performance of each architecture in terms of accuracy, precision, recall, and F1-score, both with and without data augmentation. This comparison allows us to assess the impact of data augmentation on the model's generalization capabilities.

ResNet-50 achieved an accuracy of 95.4% without data augmentation and 95.89% with augmentation, demonstrating a modest improvement. The EfficientNet V2 model showed a more significant improvement, with its accuracy increasing from 96.00% to 97.03% when data augmentation was applied. Finally, the ViT-b16 model, although still effective, showed the smallest performance gains, with its accuracy improving from 94.89% to 95.37%.

The models generally exhibited a similar performance pattern when evaluating metrics beyond accuracy, i.e., precision, recall, and F1 score. EfficientNet V2 achieved the best results overall, except for recall, where ResNet-50 performed better when trained without data augmentation. Data augmentation emerged as a crucial factor for this problem, significantly enhancing the performance of nearly all models across different scenarios.

These results indicate that, although all architectures benefit from data augmentation, EfficientNet V2 outperforms the others, both in terms of accuracy and robustness to overfitting. EfficientNet V2's superior performance can be attributed to its efficient scaling strategy, which balances network depth, width, and resolution. In contrast, the dependence of the ViT-b16 model on transformer-based architecture, which typically requires larger datasets to realize its full potential, may explain its comparatively lower performance in this study.

We also present the confusion matrices for each performed experiment in Table II. In addition to the metrics shown in Table I, confusion matrices bring details about how each model performed in classifying images from each class. Therefore, it highlights the classes that each model is good to classify or has some difficulty doing correctly.

V. Conclusion

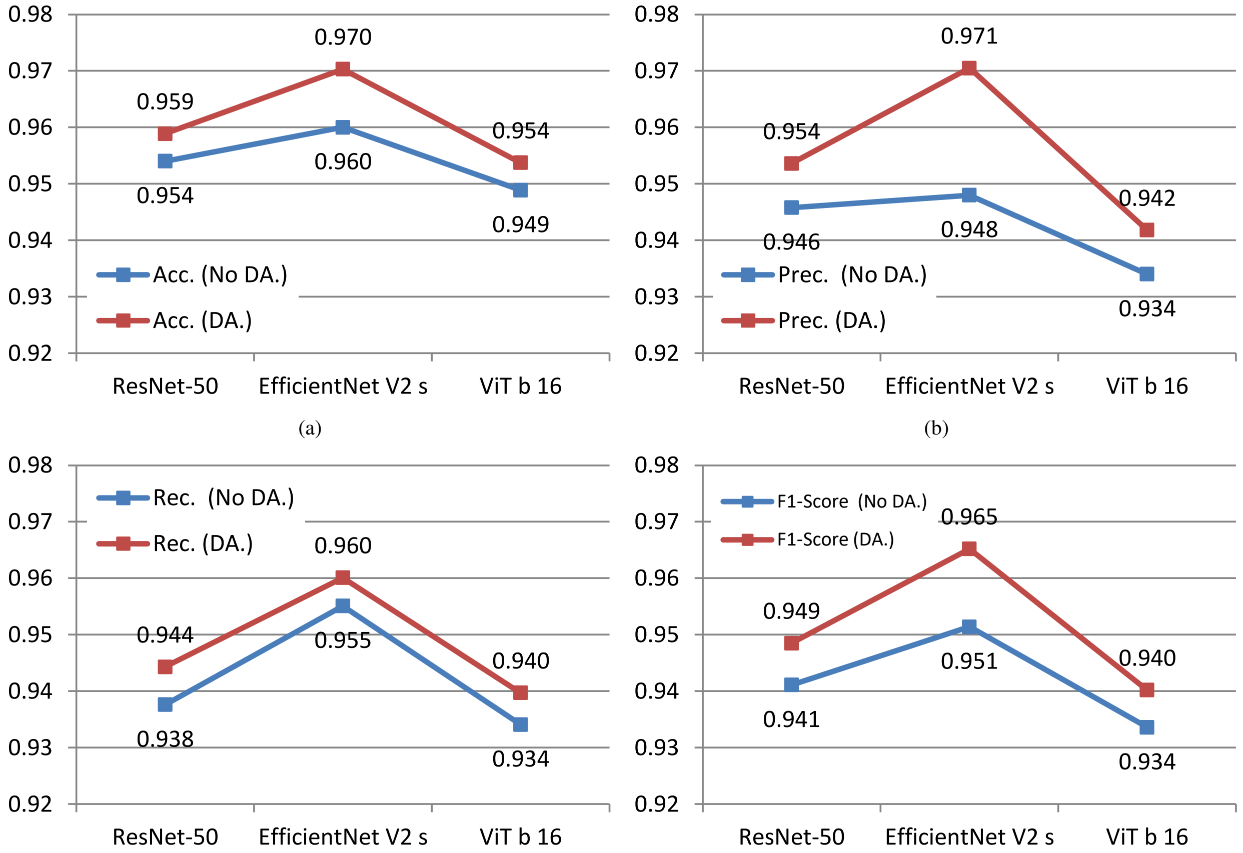
This study provides a comprehensive analysis of three deep learning architectures applied to weed species classification using the DeepWeeds dataset. Our findings highlight that EfficientNet V2, when combined with data augmentation techniques, not only outperforms other architectures in terms of accuracy but also proves robust against overfitting, delivering consistently superior performance across all metrics evaluated citetan2019efficientnet.

This combination of techniques and architecture makes EfficientNet V2 a strong candidate for deployment in real-world agricultural applications, where precision and reliability are crucial to optimizing crop management and reducing dependence on herbicides. Furthermore, the computational efficiency of EfficientNet V2, which balances model depth and complexity with available computational resources, makes it a practical solution for implementation in environments with hardware constraints, such as mobile devices or agricultural drones. Finally, applying solutions based on computer vision, such as those proposed in this study, can contribute to more efficient and sustainable agricultural practices, promoting environmentally responsible agriculture.

Future work could explore other of deep learning architectures and transfer learning to leverage pre-trained models on

TABLE I
EXPERIMENTAL RESULTS WITH MODELS TRAINED WITH AND WITHOUT DATA AUGMENTATION.

	Without Data Augmentation				With Data Augmentation			
Architecture	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1-Score**
ResNet-50	0.9540	0.9448	**0.9458**	0.9485	0.9589	0.9536	0.9443	0.9485
EfficientNet V2	**0.9600**	**0.9484**	0.9408	**0.9511**	**0.9703**	**0.9705**	**0.9601**	**0.9652**
ViT b 16	0.9489	0.9380	0.9351	0.9374	0.9537	0.9418	0.9397	0.9402



(a)

(b)

(c)

(d)

Fig. 2. Charts with the validation metrics over the test set for each model trained with and without data augmentation strategy. (a) Accuracy. (b) Precision. (c) Recall. (d) F1-score.

large related datasets, accelerating training and improving performance in weed classification. The search for more suitable data augmentation strategies would also be important, ensuring better generalization of the models. Training and evaluating these strategies with a larger set of weed species, including other datasets, would provide valuable insights to improve the accuracy and robustness of the models.

VI. ACKNOWLEDGMENTS

Omitted due to the double-blind review

REFERENCES

[1] B. G. Martha Jr, E. Contini, and E. Alves, “12. embrapa: its origins and changes,” *The regional impact of national policies: the case of Brazil*, vol. 204, 2012.

[2] E. Alves, “Embrapa: Um caso bem-sucedido de inovação institucional,” *Revista de Política Agrícola*, vol. 19, no. 2, pp. 65–73, 2010, acessado em: 9 de dezembro de 2023. [Online]. Available: https://seer.sede.embrapa.br/index.php/RPA/article/view/1115/pdf

[3] Instituto Brasileiro de Geografia e Estatística (IBGE), “PIB cresce 1,9% no primeiro trimestre de 2023, impulsionado pela Agropecuária,” 2023, accessed: 2023-12-09. [Online]. Available: https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/37030-pib-cresce-1-9-no-primeiro-trimestre-impulsionado-pela-agropecuaria

[4] A. Kubiak, A. Wolna-Maruwka, A. Niewiadomska, and A. A. Pilarska, “The problem of weed infestation of agricultural plantations vs. the assumptions of the european biodiversity strategy,” *Agronomy*, vol. 12, no. 8, 2022. [Online]. Available: https://www.mdpi.com/2073-4395/12/8/1808

TABLE II
CONFUSION MATRIX FOR THE CLASSIFICATIONS OVER THE TEST SET.

		Without Data Augmentation									With Data Augmentation								
Arch.	Classes	Chinee apple	Lantana	Negative	Parkinsonia	Parthenium	Prickly acacia	Rubber vine	Siam weed	Snake weed	Chinee apple	Lantana	Negative	Parkinsonia	Parthenium	Prickly acacia	Rubber vine	Siam weed	Snake weed
ResNet-50	**Chinese apple**	195	1	18	0	0	0	1	0	10	211	2	9	0	0	0	0	0	3
	Lantana	0	203	5	0	0	1	1	2	1	0	206	5	0	0	0	0	0	2
	Negative	7	5	1773	3	14	5	10	4	10	8	0	1789	3	2	7	4	5	3
	Parkinsonia	0	0	0	201	1	4	0	0	0	0	0	0	204	0	2	0	0	0
	Parthenium	0	2	8	1	191	2	0	0	0	0	1	7	1	195	1	0	0	0
	Prickly acacia	0	0	6	1	0	206	0	0	0	1	0	8	1	0	203	1	0	0
	Rubber vine	1	0	5	0	0	0	196	0	0	1	0	1	0	0	0	201	0	0
	Siam weed	0	0	7	0	1	0	0	207	0	0	0	9	0	0	0	0	206	0
	Snake	5	9	17	1	0	0	0	2	169	0	2	16	0	0	0	0	2	183
EfficientNet V2	**Chinese apple**	211	2	9	0	0	0	0	0	3	195	1	18	0	0	0	1	0	10
	Lantana	0	206	5	0	0	0	0	0	2	0	203	5	0	0	1	1	2	1
	Negative	8	0	1789	3	2	7	4	5	3	7	5	1773	3	14	5	10	4	10
	Parkinsonia	0	0	0	204	0	2	0	0	0	0	0	0	201	1	4	0	0	0
	Parthenium	0	1	7	1	195	1	0	0	0	0	2	8	1	191	2	0	0	0
	Prickly acacia	1	0	8	1	0	203	1	0	0	0	0	6	1	0	206	0	0	0
	Rubber vine	1	0	1	0	0	0	201	0	0	1	0	5	0	0	0	196	0	0
	Siam weed	0	0	9	0	0	0	0	206	0	0	0	7	0	1	0	0	207	0
	Snake	0	2	16	0	0	0	0	2	183	5	9	17	1	0	0	0	2	169
ViT B 16	**Chinese apple**	189	2	17	0	4	1	1	1	11	185	6	19	1	3	0	3	0	8
	Lantana	0	210	2	0	1	0	0	0	0	1	203	5	0	0	2	0	1	1
	Negative	7	9	1768	4	2	17	4	3	7	12	7	1761	3	7	12	12	6	6
	Parkinsonia	0	0	2	199	0	5	0	0	0	0	0	2	196	0	8	0	0	0
	Parthenium	2	0	8	0	193	1	0	0	0	1	0	7	0	192	3	0	0	1
	Prickly acacia	1	0	5	0	1	206	1	0	1	0	0	6	0	0	205	0	0	1
	Rubber vine	1	1	5	0	0	0	194	0	0	3	1	6	0	0	0	192	0	0
	Siam weed	1	2	8	0	0	0	0	204	0	0	2	6	0	0	0	0	207	0
	Snake	10	5	10	0	0	0	0	1	177	5	7	8	0	0	0	0	1	182

[5] B. S. Chauhan, "Grand challenges in weed management," *Frontiers in Agronomy*, vol. 1, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fagro.2019.00003/full

[6] F. Y. Adusei, M. A. Adusei, and B. Lartey, "A roadmap for sustainable disease, pest, and weed management," *Biology and Life Sciences Forum*, vol. 27, no. 1, 2023. [Online]. Available: https://www.mdpi.com/2673-9976/27/1/24

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[8] G. Mujtaba and S. W. Baik, "Weed detection using deep learning: A systematic literature review," *Sensors*, vol. 23, no. 7, p. 3670, 2023.

[9] Z. Zhang, J. Sun, M. Liu, M. Xu, Y. Wang, G.-l. Wu, H. Zhou, C. Ye, D. Tsechoe, and T. Wei, "Don't judge toxic weeds on whether they are native but on their ecological effects," *Ecology and Evolution*, vol. 10, no. 17, pp. 9014–9025, 2020.

[10] A. Olsen, D. A. Konovalov, B. Philippa, P. Ridd, J. C. Wood, J. Johns, W. Banks, B. Girgenti, O. Kenny, J. Whinney *et al.*, "Deepweeds: A multiclass weed species image dataset for deep learning," *Scientific reports*, vol. 9, no. 1, p. 2058, 2019.

[11] M. H. Saleem, K. K. Velayudhan, J. Potgieter, and K. M. Arif, "Weed identification by single-stage and two-stage neural networks: A study on the impact of image resizers and weights optimization algorithms," *Frontiers in Plant Science*, vol. 13, p. 850666, 2022.

[12] Y. Yang, Y. Li, J. Yang, and J. Wen, "Dissimilarity-based active learning for embedded weed identification," *Turkish Journal of Agriculture and Forestry*, vol. 46, no. 3, pp. 390–401, 2022.

[13] Y. Wang, S. Zhang, B. Dai, S. Yang, and H. Song, "Fine-grained weed recognition using swin transformer and two-stage transfer learning," *Frontiers in Plant Science*, vol. 14, p. 1134932, 2023.

[14] Y. Wu, Y. He, and Y. Wang, "Multi-class weed recognition using hybrid cnn-svm classifier," *Sensors*, vol. 23, no. 16, p. 7153, 2023.

[15] A. M. Hasan, D. Diepeveen, H. Laga, M. G. Jones, and F. Sohel, "Image patch-based deep learning approach for crop and weed recognition," *Ecological informatics*, vol. 78, p. 102361, 2023.

[16] N. Belissent, J. M. Peña, G. A. Mesías-Ruiz, J. Shawe-Taylor, and M. Pérez-Ortiz, "Transfer and zero-shot learning for scalable weed detection and classification in UAV images," *Knowledge-Based Systems*, vol. 292, p. 111586, May 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705124002211

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005622000911

[21] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.