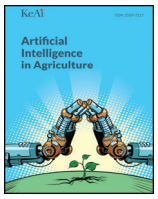
Artificial Intelligence in Agriculture 6 (2022) 242–256

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



A study on deep learning algorithm performance on weed and crop species identification under different image background

Sunil G C [a], Cengiz Koparan [a], Mohammed Raju Ahmed [a], Yu Zhang [a], Kirk Howatt [b], Xin Sun [a,*]

[a] Department of Agricultural and Biosystems Engineering, North Dakota State University, Fargo, ND 58108, USA
[b] Department of Plant Sciences, North Dakota State University, PO Box 6050, Fargo, ND 58108-6050, USA

ARTICLE INFO

Article history:
Received 26 September 2022
Received in revised form 8 November 2022
Accepted 9 November 2022
Available online 12 November 2022

Keywords:
Computer vision
Deep learning
Image background
Neural network
Weed classification

ABSTRACT

Weed identification is fundamental toward developing a deep learning-based weed control system. Deep learning algorithms assist to build a weed detection model by using weed and crop images. The dynamic environmental conditions such as ambient lighting, moving cameras, or varying image backgrounds could affect the performance of deep learning algorithms. There are limited studies on how the different image backgrounds would impact the deep learning algorithms for weed identification. The objective of this research was to test deep learning weed identification model performance in images with potting mix (non-uniform) and black pebbled (uniform) backgrounds interchangeably. The weed and crop images were acquired by four canon digital cameras in the greenhouse with both uniform and non-uniform background conditions. A Convolutional Neural Network (CNN), Visual Group Geometry (VGG16), and Residual Network (ResNet50) deep learning architectures were used to build weed classification models. The model built from uniform background images was tested on images with a non-uniform background, as well as model built from non-uniform background images was tested on images with uniform background. Results showed that the VGG16 and ResNet50 models built from non-uniform background images were evaluated on the uniform background, achieving models' performance with an average f1-score of 82.75% and 75%, respectively. Conversely, the VGG16 and ResNet50 models built from uniform background images were evaluated on the non-uniform background images, achieving models' performance with an average f1-score of 77.5% and 68.4% respectively. Both the VGG16 and ResNet50 models' performances were improved with average f1-score values between 92% and 99% when both uniform and non-uniform background images were used to build the model. It appears that the model performances are reduced when they are tested with images that have different object background than the ones used for building the model.

1. Introduction

One of the major causes of yield loss in agriculture is due to weeds. Oerke (2006) reported that weeds (34%) caused approximately double yield losses than that of animal pests (18%), and pathogens (16%) (Oerke, 2005). A recent study by Soltani et al. (2017) estimated that soybean yield would be reduced due to weed interference by approximately 52%, which is equivalent to $16 billion annually in the US. Besides yield loss, weeds make harvesting difficult, give undesired color and test on crops, and act as hosts for disease and pests (Holzner, 1982). Hence, a number of studies have been performed to develop a precision weed control system by the use of computer vision-based deep learning architecture, and robotics, which has been depicted in a recent review article by Li et al. (2022). Deep learning is the subfield of machine learning, which is inspired by the artificial neural network algorithm. It is one of the breakthroughs in the area of computer vision for image classification, object detection, and localization, which has been effectively used in weed detection for robotic weed control research (Hasan et al., 2021).

Deep learning architectures and algorithms were able to achieve near 98% performance accuracy in weeds classification (Espejo-Garcia et al., 2020; Hu et al., 2020). However, most of the research is location-specific, crop-specific, and weed-specific (Espejo-Garcia et al., 2020; Hasan et al., 2021; Khan et al., 2021; Olsen et al., 2019). The deep learning model built from a certain section of the weed and crop images may not be able to perform well in another section due to the different soil types, lightning condition, or image background situations. However, there is limited study on weed classification deep learning model evaluation on completely unseen nature of images with completely different image backgrounds in the area of sites-specific precision weed control images lacking generality and robustness (Espejo-Garcia et al., 2020; Wu et al., 2021). Such type of

* Corresponding author.
E-mail address: xin.sun@ndsu.edu (X. Sun).

https://doi.org/10.1016/j.aiia.2022.11.001

study is important because, during the deep learning computer vision model training, crop and weeds pixels along with soil background pixels are used by convolutional neural network architecture to generalize the weed detection model (Kulawardhana, 2011). The changes in image pixels between the model training images and testing images might lead to the model performance degradation (Velumani et al., 2021).

A recent study cross evaluated the deep learning model in different background condition in disease detection area of agriculture (Ferentinos, 2018). In a previous study by Ferentinos (2018), model built from labrotory condition images was tested on images obtained in field, which caused decline in model performance due to having completely different background image pixel than that of labrotory images.There is higher probability of having different background in disease detection studies images than weed detection images, because weed detection images capture whole plant images whereas, disease detection images capture s specific region of plant affected by disease. Similar studies are important to fill the existing gap on research about effects of background in weed identification which could assist toward the development of robust real time deep learning weed detection model in agriculture industries.

The objective of this study was 1) to build comprehensive weed classification models using deep learning algorithms on weed species of Horseweed, Palmer Amaranth, Redroot Pigweed, Kochia, Ragweed, and Waterhemp among crop species of Canola and Sugar beet; 2) to evaluate the deep learning model performance on two different backgrounds of the acquired weed and crop image pictures in greenhouse condition.

2. Material and methods

2.1. Weed and crop greenhouse layout design, image acquisition, and image labeling

The crops and weeds were grown in 3.5 cubic inch size pots in a greenhouse environment for indoor data collection. The weed seeds of Horseweed, Palmer Amaranth, Redroot Pigweed, Ragweed, Waterhemp, and Kochia and crop species of Canola and Sugar beets were planted in the pots for weed identification study. The pots were randomly placed on a bench. The selected weed and crop species are prominent weed species and major crops of the Midwest region of the United States (Bryson and DeFelice, 2010). To create a two different image background scenario, the black lava pebble gravel material with an average of 2-5 mm diameter was placed at the surface of the potting mix approximately a 6 mm thick layer to ensure proper surface coverage. The average weight of the gravel material was 150 g for each square pot. Fig. 1 depict the weed and crop pots with potting mix and gravel material to create background scenario-1 (non-uniform) in Fig. 1a and background scenario-2 (uniform) in Fig. 1b, respectively.

Four Canon EOS T7 digital cameras (Canon Inc., Tokyo, Japan) were mounted over the table in a fixed position at a height (≈48 ± 0.5 in.), where camera captures the total width of the bench. The reason for using four cameras was to cover entire bench area (≈48.8 square foot). The cloud based automatic image acquisition system was used to acquire weed and crop images. Images were captured for three days with non-uniform background and uniform background to minimize the error that could cause due to the plant growth. The images were captured within three days of 30 March 2021 with natural lighting during the day and artificial lighting during the evening in the greenhouse.

After images were acquired, the next step was extracting the individual weed and crop species from the single image. A Python (Python Software Foundation, Wilmington, DE) script was developed to label a single image of a day using an OpenCV library (Bradski, 2000), which saves the bounding box coordinate in order to perform automatic cropping of the remaining images for the same day. Before applying the python script, images were manually checked if there were any disturbances, such as human intervention, external objects in the image, and changed illumination. The images that included disturbances

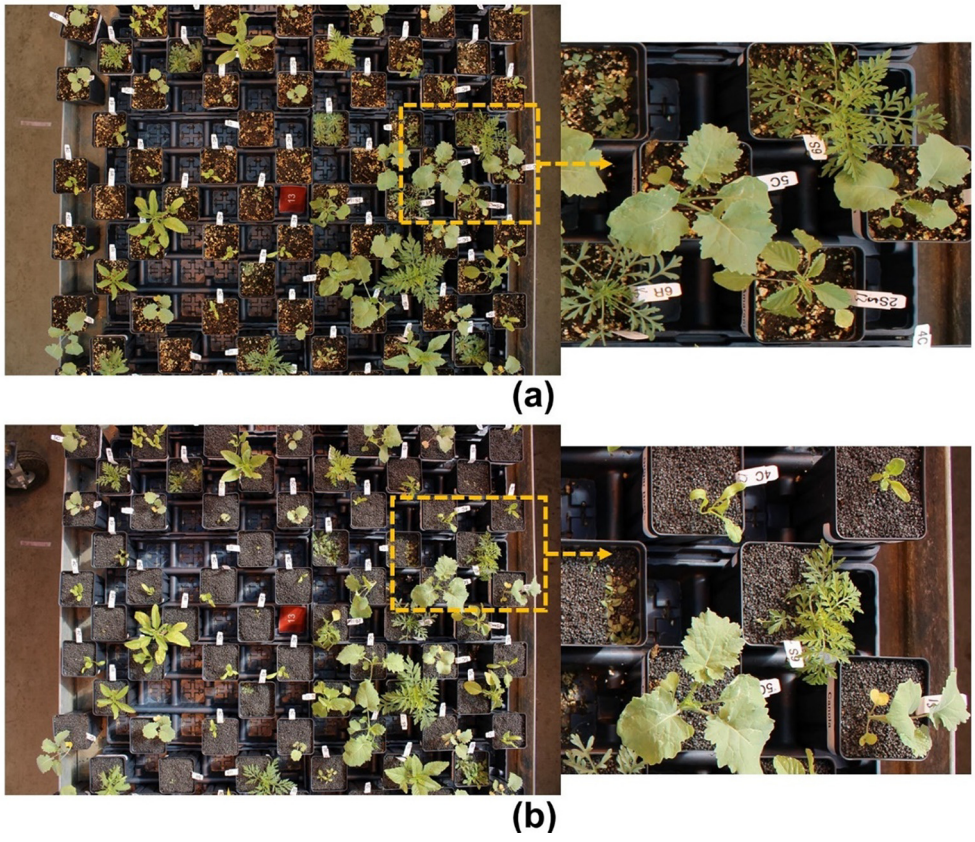


Fig. 1. Image captured; (a) before adding gravel (non-uniform background), and (b) after adding the gravel (uniform background).

were removed before processing the dataset. Fig. 2 shows a detailed flowchart of image labeling and cropping python script.

After cropping the images for all the species, images were checked if any plant changed its position or only a small part within the images; such images were removed from the datasets. Fig. 3 shows the cropped images for all the species of weeds and crops for both non-uniform background and uniform background. The cropped images were labeled by making a directory for each species and placing crop images into the respective directory by the python script. Two directories were created for the 2868 images with non-uniform background datasets and 3488 images with uniform background datasets. Table 1 shows the detail of the number of each species images for both background scenario data.

2.2. Crops and weed image classification using convolutional neural network

In this study, Convolutional Neural Networks (CNN) based Visual Geometry Group 16 (VGG16) deep learning architecture was used to compare the classification of the weed and crop images for the non -uniform background scenario (S1) and uniform background scenario (S2). VGG16 architecture already trained on ImageNet 1000 classes was used to finetune the model with weed and crop image datasets (Abdalla et al., 2019; Espejo-Garcia et al., 2020). VGG16 model was trained up to 50 epochs with 32 steps in each epoch. This technique is called transfer learning, where convolution and pooling layers were frozen and fully connected layers were modified for new sets of problems. Instead of training and optimizing frozen layers, weights and biases, weight, and biases from previously trained VGG16 models on ImageNet datasets were used. Only the fully connected layer was trained with eight classes as the output classes. Transfer learning was used because it makes training faster and is able to achieve outstanding performance with smaller numbers of data (GC et al., 2021). Fig. 4a shows the complete block diagram for VGG16 architecture model layers used in this study with their outputs and input dimension (Simonyan and Zisserman, 2014). The complete diagram depicts the ResNet50 architecture with its trainable and frozen layers for the transfer learning approach (He et al., 2015). There were five blocks of convolution and pooling layers that were frozen during the training. In VGG16 model, there were a total of 14,915,400 parameters, out of which 200,712 parameters were trainable, and the remaining 14,714,688 parameters were not trainable because only few outermost layers were trained in transfer learning model training approach.

Similar to VGG16, CNN based Residual Networks 50 (ResNet50) deep learning architecture was also used to train the non-uniform background scenario (S1) data, uniform background scenario (S2) data, and combined-datasets scenario (C) obtained after merging both scenarios' data. For this training, a ResNet50 model already trained on the DeepWeeds image dataset by Olsen et al. (2019) was used. A transfer learning technique was used for ResNet50, where only 20,488 parameters were trained, and the remaining 23,583,616 parameters were frozen. ResNet50 was trained for 70 epochs, with 32 steps in each epoch. An outermost layer with 9 output classes and sigmoid activation function was removed and a new outer layer was added for 8 output classes a with softmax activation function. Softmax assigns the decimal

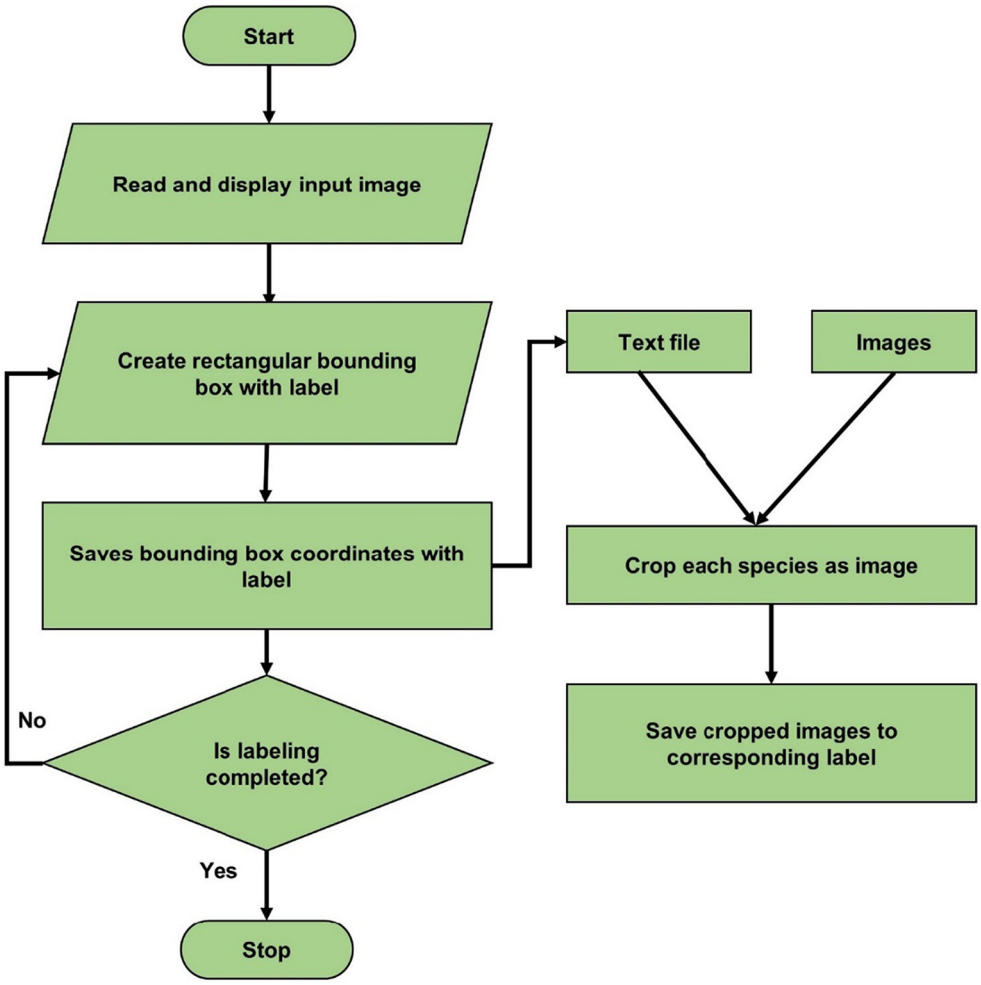


Fig. 2. Flowchart showing labeling of each image for weeds and crops species from the single image captured by the camera.

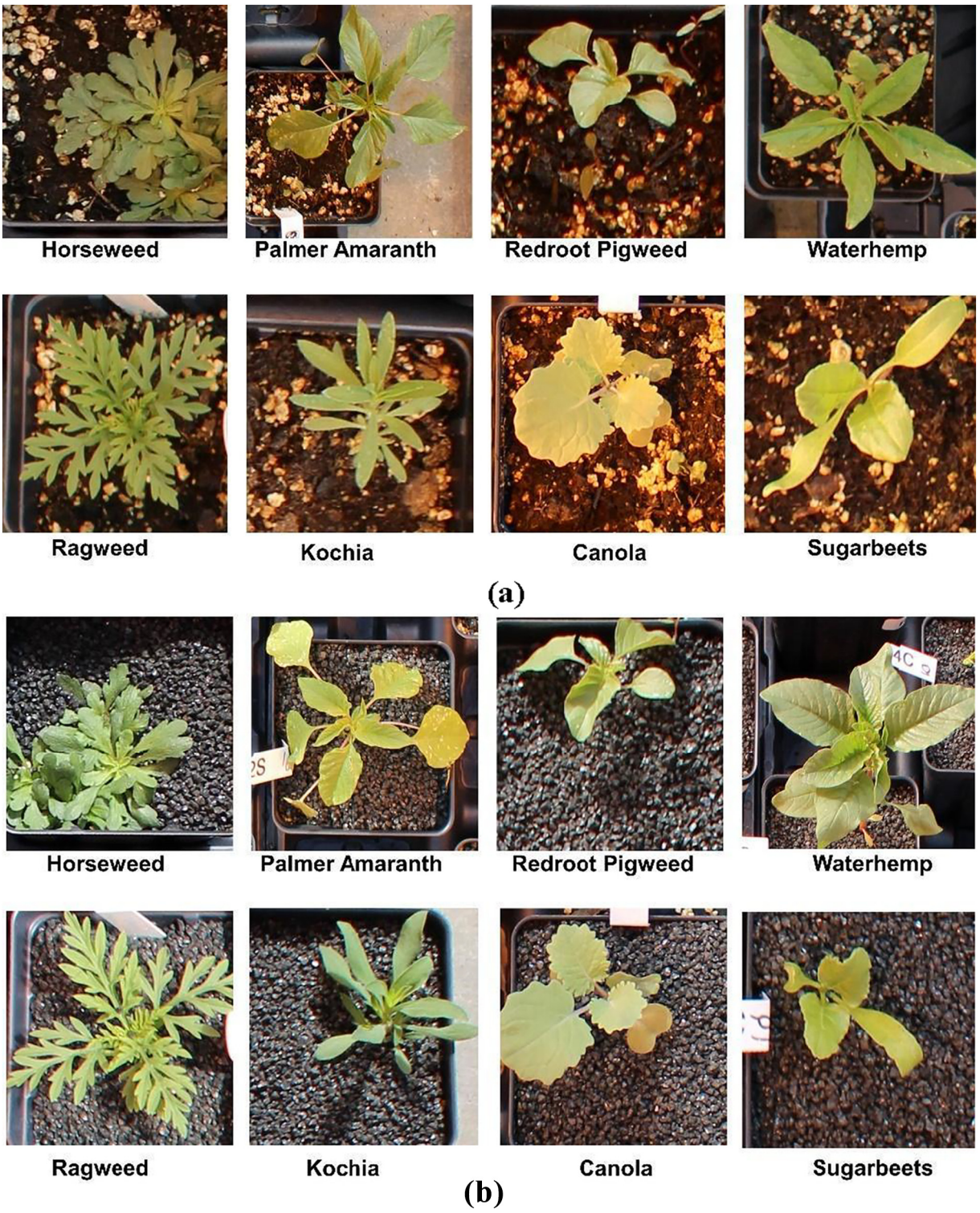


Fig. 3. Copped images of crop and weed plants from the image captured by canon camera after the application of labeling and cropping python scripts (a). Weed and crop images before the application of black gravel (non-uniform background) (b). Weeds and crop images after the application of black gravel (uniform background).

probabilities to each class of weeds and crops, where the class with the highest probability was considered to be the predicted class (Alzubaidi et al., 2021).

VGG16 and ResNet50 model training and validation were performed on both S1, S2, and C datasets. Labeled images were randomly divided into the training, validation, and testing datasets for both scenario data in the ratio of 60%, 20%, and 20% respectively. There were 1719 and 1149 number of crop and weed images for S1, respectively. Whereas, S2 datasets had 1984 and 1504 number of crop and weed images respectively. The more details on number of images for individual weed and crop species is briefly shown in Table 1. Firstly, a model with S1 and S2 datasets was developed. Later, both training and validation datasets were mixed, and a third model was developed with combined data to create C. A detailed flowchart for the steps of VGG16 and ResNet50

Table 1
The total number of training, validation, and testing images of crop and weed species for non-uniform background scenario (S1) and uniform background scenario (S2) data.

	Training		Validation		Testing		Total		Grand total
	S1	S2	S1	S2	S1	S2	S1	S2	Both
Horseweed	132	144	45	48	45	48	222	240	462
Palmer Amaranth	124	180	42	60	42	60	208	300	508
Redroot Pigweed	62	87	22	30	22	30	106	147	253
Ragweed	140	191	47	65	47	65	234	321	555
Waterhemp	134	182	45	62	45	62	224	306	530
Canola	522	640	175	214	175	214	872	1068	1940
Kochia	93	114	31	38	31	38	155	190	340
Sugar beets	507	548	170	184	170	184	847	916	1763
Total	**1714**	**2086**	**577**	**701**	**577**	**701**	**2868**	**3488**	**6356**

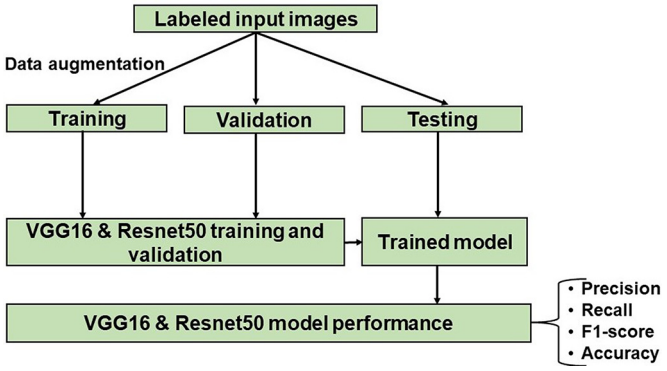


Fig. 5. Diagram showing VGG16 and ResNet50 model development flowchart for training, validation, and testing steps.

model development, validation, and testing steps is depicted in Fig. 5. Adam optimizer (Kingma and Ba, 2015) and categorical cross-entropy were used as optimizer and cost function respectively for both VGG16 and ResNet50 model training and validation. Training datasets were only augmented to increase the number of datasets, which also helps to reduce the model overfitting when training with models with larger numbers of parameters. Image augmentation techniques used were re-scale, shear, shift, flip, rotation, and zoom. Finally, the models developed from training and validation datasets were tested with the testing datasets to evaluate the performance of the models. Both models were built and tested by using deep learning TensorFlow (TensorFlow.org) and Keras (keras.io) python Application Program Interface (API). The training and testing were performed on desktop computer with Intel® Core™ i7–3770 CPU @ 3.40 GHz processor, 8.00 GB RAM memory (Dell 7040, Dell Technologies Inc., Round Rock, TX).

2.3. Crops and weeds data analysis

In this study, to see the effects of background scenarios on model performance, models were tested with images with different background than the images used for building the model. Test datasets from both S1 and S2 were tested on three models trained and validated with S1 model, S2 model, and C model datasets. This helped to cross-check the model performance when the test dataset's background scenario was completely different from the scenario on which the model was developed.

VGG16 and ResNet50 model performance was measured with precision, recall (sensitivity), f1-score, and accuracy parameters. The f1-score metric was used over accuracy because it works well when there is an imbalance in the number of images in each class (Johnson and Khoshgoftaar, 2019). Accuracy is the number of correctly classified

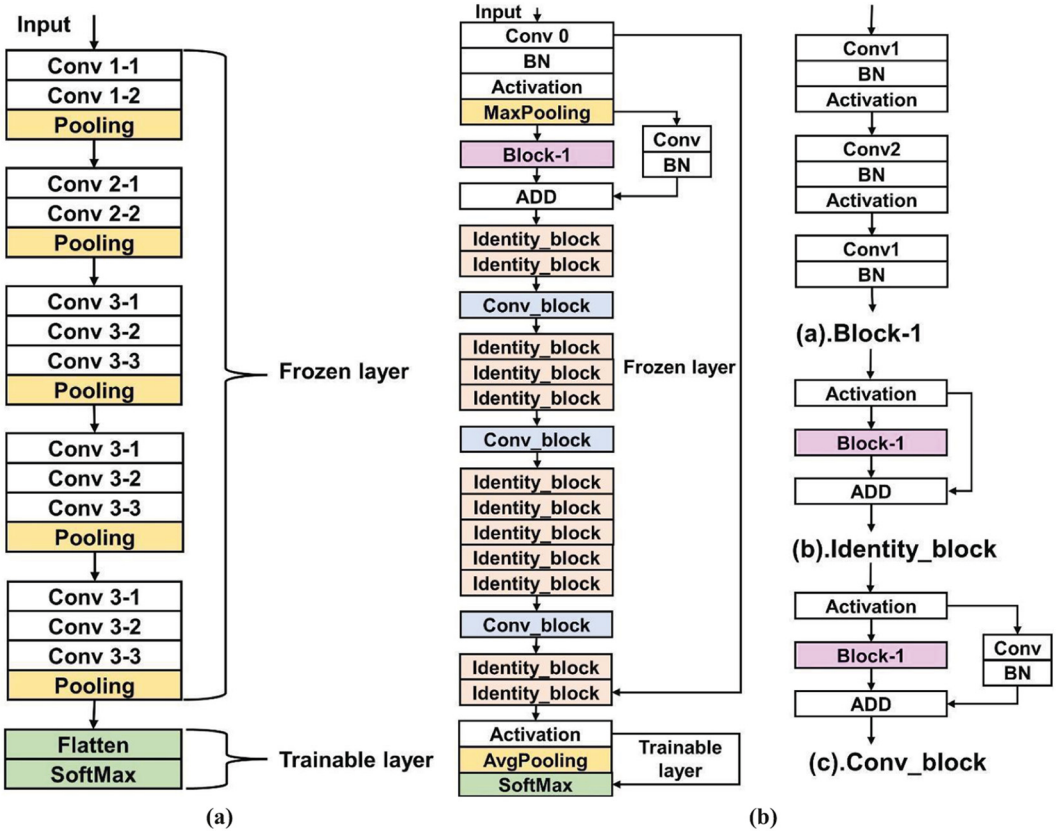


Fig. 4. VGG16 and ResNet50 architecture showing input layer, convolution layer, pooling layer, and output layer. Frozen layers parameters are not trained, and trainable layer parameters are optimized during the training. (a). VGG16 architecture (b). ResNet50 architecture.

data instances over the total number of data instances. Accuracy may not be a good measure when datasets are imbalanced because it might not show a clear picture of model performance. However, the f1-score, which is also the geometric mean of the precision and recall depict clear pictures of model performance in terms of each class. Precision is the percentage of correctly predicted images out of the total number of predicted images, whereas recall is the percentage of correctly predicted images out of the total number of images in the actual class. Ideally, for good classifiers, both precision and recall values should be 1, but they counter each other, and increasing one usually reduces the other (Arya et al., 2020). The macro and weighted average were calculated for the precision, recall, and f1-score. A confusion matrix was also used to visualize the summary of the prediction results of the model. For calculating all the metrics and visualizing the confusion matrix, python sklearn API was used (Pedregosa et al., 2011).

The decline in model performance for S1 and S2 was obtained by calculating the percentage of error (PE). It was calculated by finding the difference between the f1-score or accuracy value of the model when it was tested on the same background scenario test datasets and the different background scenario test data sets. Furthermore, the percentage of improvement (PI) was calculated for both background scenario models when the models were built with combined datasets. The significance of the percentage of error PE and PI was also tested with a one-tailed paired *t*-test.

3. Results

3.1. The model performances from training and validation steps for crop and weed classification

The model performance evaluation of VGG16 and ResNet50 model has been made for S1, S2, and C datasets during the model training and validation to check the model generalization. The training and validation accuracy of VGG16 model was 97% for both S1 and S2 data. Fig. 6 shows the training accuracy, validation accuracy, training loss, and validation loss from the VGG16 and ResNet50 model training. When the epoch was increased from epoch 1 to 50 as training continued, training loss decreased from 1.6501 to 0.0469, training accuracy increased from 54.42% to 97.85%, validation loss decreased from 0.6675 to from 0.0176, and validation accuracy increased from 75% to 99.02%. It appeared that the accuracy was increased, and loss was decreased for VGG16-S1 model (Fig. 6). A similar trend was found for the VGG16-S2 model. For the VGG16-S2 model training (Fig. 6), training loss decreased from 1.8856 to 0.0262, training accuracy increased from 49.02% to 99.22%, validation loss decreased from 0.6767 to 0.0562; validation accuracy increased from 79.10% to 97.85%. This showed validation accuracy was higher for the VGG16-S1 model than the VGG16-S2 model.

Similar to the VGG16, training and validation accuracy was increased, and training and validation loss was decreased for ResNet50 model training for both background scenarios, when epoch was increased from 1 to 70 (Fig. 6). During the ResNet50-S1 model training, training loss decreased from 1.3279 to 0.0698, training accuracy increased from 57.62% to 98.63%; validation loss decreased from 0.9317 to 0.1018, and validation accuracy increased from 68.36% to 97.27%. Whereas, during the ResNet50-S2 model training, training loss decreased from 1.3578 to 0.1217, training accuracy increased from 55.27% to 97.57%,validation loss decreased from 0.9924 to 0.1372, and validation accuracy increased from 64.65% to 95.31%.This showed both S1 and S2 ResNet50 models achieved a validation accuracy of greater than 95%. However, the ResNet50-S1 model achieved higher validation accuracy than the ResNet50-S2 model.

Combined datasets from both S1 and S2 were used to train and validate the VGG16 and ResNet50 models. During the VGG16 model training on VGG16-C model (Cmodel) from epoch 1 to 50, as training continued accuracy and loss values changed as shown in Fig. 6. Training loss decreased from 1.9313 to 0.0873, training accuracy increased from 51.95% to 97.66%, validation loss decreased from 1.0262 to 0.0339, and validation accuracy increased from 70.51% to 98.24%. Similarly, for ResNet50-C model development, as the epoch increased from 1 to 70, training loss decreased from 1.2242 to 0.1380, training accuracy increased from 56.84% to 95.31%, validation loss decreased from 1.0244 to 0.1563, and validation accuracy increased from 64.26% to 93.55%. Training and validation accuracy for both VGG16-C and ResNet50-C models were obtained above 95% (ResNet50: epoch 66). VGG16-C model training and validation accuracy were higher than ResNet50-C model for weed, and crop datasets used in this study. All six training graphs in Fig. 6 shows model was not overfitted during the training steps.

3.2. The VGG16 model prediction performance on test images for weed and crop classification

For deep learning model performance evaluation, precision, recall, f1-score, and confusion matrix were used to depict the VGG16 model performance on S1 test data (S1tedata) and S2 test data (S2tedata). The VGG16-S1, VGG16-S2, and VGG16-C models were tested on both S1tedata and S2tedata. Table 2 shows the performance of the VGG16-S1 model in terms of precision, recall, and f1-score. Model performance was the highest with an average f1-score of 99% when the VGG16-S1 model was tested with S1tedata. This result would be expected because the model learned from images with similar backgrounds. However, when the VGG16-S1 model was tested with S2tedata, the macro average f1-score declined to 83% and the weighted average f1-score declined to 87%. Among two crops and six weed species, the top three species for which f1-score severely declined were Redroot Pigweed (100% to 46%), Horseweed (100% to 79%), and Palmer Amaranth (99% to 80%). Similar type of performance degradation was observed by Ferentinos (2018) in plant disease detection when test images were changed, where model success rate was 32% to 66% when model trained on laboratory and field data was tested with field and laboratory data, respectively. The decline in the f1-score value can be attributed to the low recall value, which was due to the classification of a smaller number of Redroot Pigweed images correctly out of the total number of images in the actual class.

Fig. 7a and b shows the confusion matrix for the VGG16-S1 model, which underpin the results in Table 2. When VGG16-S1 model was tested with S1tedata, 100% of images were predicted correctly for all classes except Sugar beet with 96.5% accuracy. However, when VGG16-S1 model was tested with S2tedata, prediction accuracy declined from 98.96% to 88% due to a sharp decline in the prediction of Redroot Pigweed (100% to 33%) and Horseweed (100% to 66.7%), which is clearly depicted in Fig. 7b. While investigating the S1modelS2tedatad confusion matrix, 14.6% of Horseweed was misclassified as Ragweed and Canola. Whereas 33% and 20% of Redroot Pigweed were misclassified as Canola and Sugar beet, respectively. These misclassifications contributed in VGG16-S1 model performance degradation when the model was tested with S2tedata. The Redroot Pigweed was not promising, which may be due to the different background scenarios and lower number of training images. The 10% of Redroot Pigweed was misclassified as Palmer Amaranth, which may be due to the similar morphological characteristics of their leaves (Ma et al., 2015).

The VGG16-S2 model was tested with S1tedata and S2tedata of which performance is depicted in Table 3. The VGG16-S2 model performance seems promising with macro and weighted average f1-score value of 98% and 99%, respectively for S2tedata. For S2tedata testing, out of eight classes, Waterhemp had the lowest f1-score value of 95% whereas Canola and Ragweed had the highest f1-score value of 100%. While investigating the model performance on confusion matrix (Fig. 7c) for S2tedata, all the classes were predicted 100% correctly except Redroot Pigweed, Kochia, and Sugar beet with 93.3%, 97.4%, and 96.2% accuracies, respectively. However, the VGG16-S1 model did not

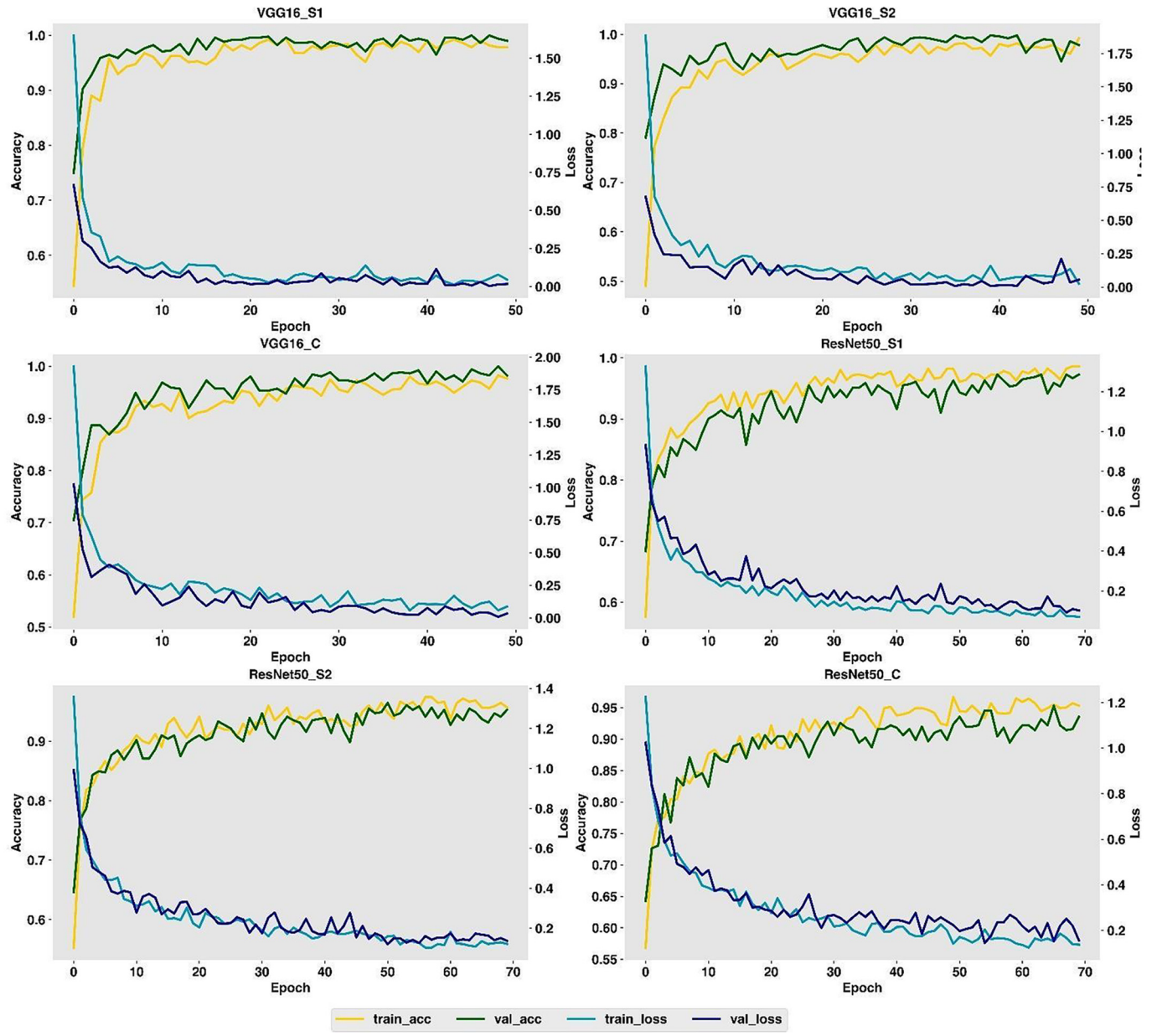


Fig. 6. Training and validation progress during the VGG16 and ResNet50 model development showing training and validation accuracy (train_acc, valid_acc) and loss (train_loss,val_loss). VGG16_S1 is the VGG16 model trained on background scenario-1 (S1) data, VGG16_S2 is the VGG16 model trained on background scenario-2 (S2) data, VGG16_C is the VGG16 model trained on combined (C) data from both background scenarios, ResNet50_S1 is the ResNet50 model trained on S1 data, ResNet50_S2 is the ResNet50 model trained on S2 data, and ResNet50_C is the ResNet50 model trained on combined (C) data from both background scenarios.

Table 2
VGG16-S1 model performance in terms of precision, recall, and f1-score metrics from the test on S1tedata (**S1modelS1tedata**) and S2tedata (**S1modelS2tedata**).

	precision		recall		f1-score		number of test data	
	S1modelS1tedata	S1modelS2tedata	S1modelS1tedata	S1modelS2tedata	S1modelS1tedata	S1modelS2tedata	S1modelS1tedata	S1modelS2tedata
Horseweed	1.00	0.97	1.00	0.67	1.00	0.79	45	48
Palmer Amaranth	0.98	0.77	1.00	0.83	0.99	0.80	42	60
Redroot Pigweed	1.00	1.00	1.00	0.30	1.00	0.46	22	30
Ragweed	1.00	0.86	1.00	1.00	1.00	0.92	47	65
Waterhemp	1.00	0.93	1.00	0.90	1.00	0.92	45	62
Canola	0.97	0.88	1.00	0.98	0.99	0.93	175	214
Kochia	1.00	0.94	1.00	0.89	1.00	0.92	31	38
Sugar beet	1.00	0.88	0.96	0.88	0.98	0.88	170	184
macro avg	**0.99**	**0.90**	**1.00**	**0.81**	**0.99**	**0.83**	577	**701**
weighted avg	**0.99**	**0.89**	**0.99**	**0.88**	**0.99**	**0.87**	577	**701**

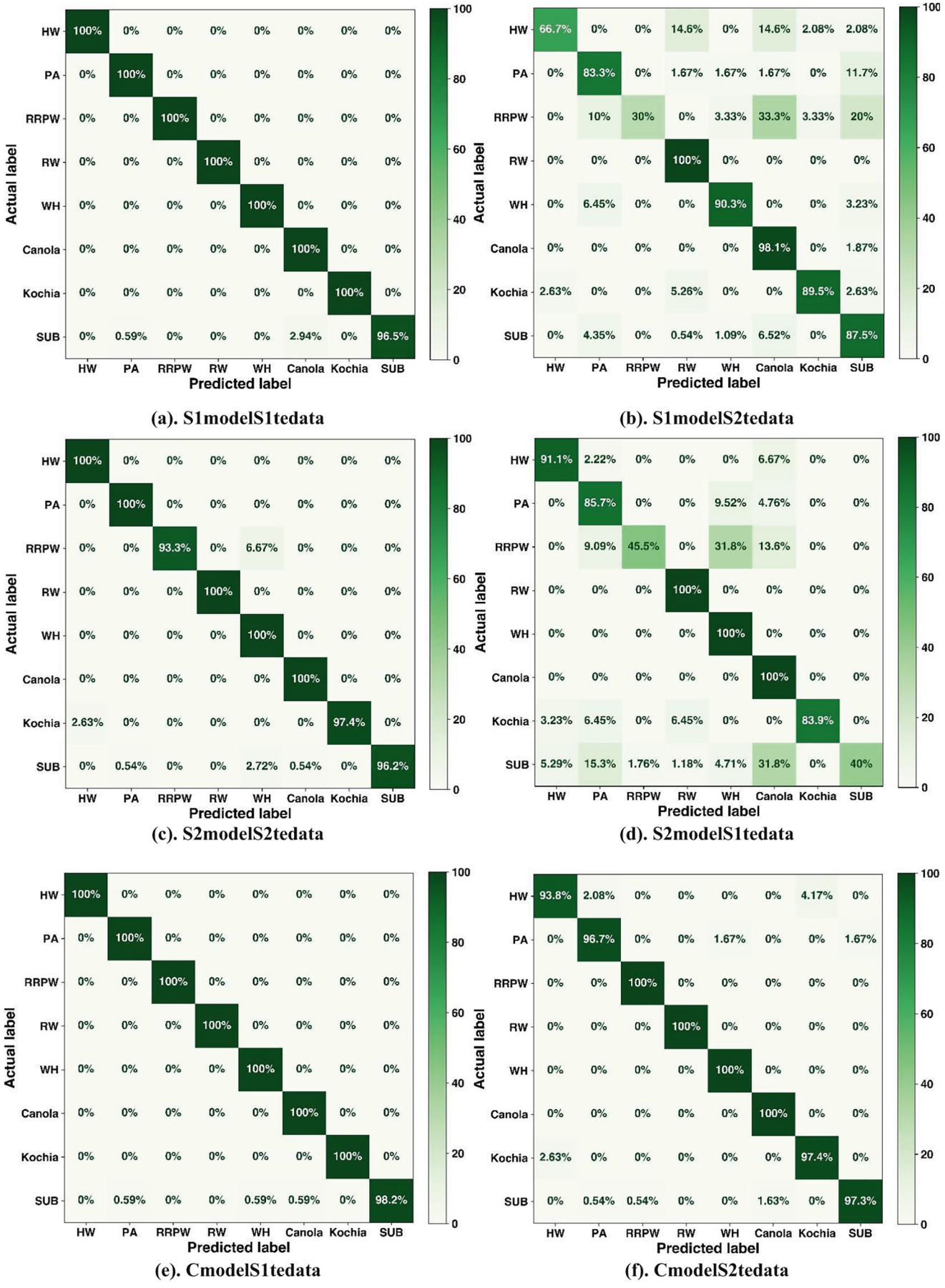


Fig. 7. VGG16 model test results in terms of the confusion matrix. (a). S1model tested with S1tedata (b). S1model tested with S2tedata (c). S2model tested with S2tedata (d). S2model tested with S1tedata (e). Cmodel tested with S1tedata (f). Cmodel tested with S2tedata (HW: Horseweed, PA: Palmer Amaranth, RRPW: Redroot Pigweed, RW: Ragweed, WH: Waterhemp, SUB: Sugar beet).

Table 3
The VGG16-S2 model performance in terms of precision, recall, and f1-score metrics for S1tedata (**S2modelS1tedata**) and S2tedata (**S2modelS2tedata**).

	precision		recall		f1-score		number of test data	
	S2modelS2tedata	S2modelS1tedata	S2modelS2tedata	S2modelS1tedata	S2modelS2tedata	S2modelS1tedata	S2modelS2tedata	S2modelS1tedata
Horseweed	0.98	0.80	1.00	0.91	0.99	0.85	48	45
Palmer Amaranth	0.98	0.54	1.00	0.86	0.99	0.66	60	42
Redroot Pigweed	1.00	0.77	0.93	0.45	0.97	0.57	30	22
Ragweed	1.00	0.92	1.00	1.00	1.00	0.96	65	47
Waterhemp	0.90	0.70	1.00	1.00	0.95	0.83	62	45
Canola	1.00	0.74	1.00	1.00	1.00	0.85	214	175
Kochia	1.00	1.00	0.97	0.84	0.99	0.91	38	31
Sugar beet	1.00	1.00	0.96	0.40	0.98	0.57	184	170
macro avg	0.98	**0.81**	**0.98**	**0.81**	**0.98**	**0.78**	**701**	577
weighted avg	**0.99**	**0.83**	**0.99**	**0.78**	**0.99**	**0.75**	**701**	577

perform well for S1tedata with macro and weighted average f1-score value of 78% and 75%, respectively. For S2modelS1tedata testing, Redroot Pigweed, Sugar beet, and Palmer Amaranth had the lowest f1 score value of 57%, 57%, and 66% respectively. F1-score value was lower in Redroot Pigweed and Sugar beet due to low recall values of 45% and 40%, respectively; however, Palmer Amaranth had lower f1-score value due to low precision value of 54%.

Moreover, while looking into the confusion matrix for S2modelS1tedata (Fig. 7d), the prediction accuracy decreased severely to 45.5% for Redroot Pigweed and 40% for Sugar beet. The 31.8% of all the Redroot Pigweed and Sugar beet images (31.8%) were misclassified as Waterhemp and Canola, respectively. This may be due to the different test images background scenario, a low number of Redroot Pigweed images, and similar morphology between the Redroot Pigweed and Waterhemp.

The VGG16-C model was tested with S1tedata and S2tedata of which the model performance is clearly depicted in Table 4. The VGG16-C model performance seems promising with both macro average and weighted average f1-score values over 98% for both S1tedata and S2tedata. For S1tedata, f1-score was either 99% or 100% for eight classes, however, for S2tedata, the f1-score value was between 96% (Kochia) and 100% (Ragweed). Moreover, Fig. 7e and Fig. 7f show the confusion matrix for CmodelS1tedata and CmodelS2tedata testing. For S1tedata, all the classes were predicted 100% correctly except Sugar beet (98.2%), whereas for S2tedata, Redroot Pigweed, Ragweed, Waterhemp, and Canola were predicted 100% correctly. The lowest prediction percentage of the S2tedata was for Horseweed images (94%). Model performance was improved when combined datasets were used to train and validate the model. Researchers have suggested using a larger dataset with a large variety of images while training the deep learning model (Hasan et al., 2021).

3.3. The Resnet50 model prediction performance on test images for weed and crop classification

The ResNet50 model was also used in this study to see if the model performance under different background conditions follows the similar pattern as that of VGG16. The model trained from S1 data (S1model) was tested with both S1 data (S1tedata) and S2 data (S2tedata). Table 5 shows the precision, recall, and f1-score for ResNet50-S1model tested on both S1tedata and S2tedata. When the ResNet50-S1 model was tested with S1tedata, the model performed well with a weighted and macro average f1-score of 98%. However, when ResNet50-S1 model was tested with S2tedata, model performance declined sharply with macro and weighted average f1-score of 75% and 80%. While drilling down into Table 5, S1modelS1tedata had f1-score of 100% for Horseweed, Ragweed, and Kochia, however none of S1modelS2tedata classes had 100% f1-score. S1modelS1tedata Redroot Pigweed had the lowest f1-score value of 93%.

In S1modelS2tedata, Ragweed had the highest f1-score of 96% and Palmer Amaranth had the lowest f1-score of 44%, where the bottom-3 classes with the lowest f1-score were Palmer Amaranth (44%), Redroot Pigweed (49%), and Waterhemp (75%). Palmer Amaranth had 94% precision but 28% recall value, which had caused severe decline in f1-score. However, Redroot Pigweed had both lower precision (63%) and recall (40%) value causing sharp decline in f1-score in S2tedata from that of S1tedata. The pattern of degradation in the model performance is similar for both VGG16 and ResNet50 models when the background of the test data was different from the background of the trained model data.

The clearer picture of the ResNet50-S1 model performance on S1tedata and S2tedata is depicted in Fig. 8a and b confusion matrix. For S1modelS1tedata, the model predicted Horseweed, Palmer Amaranth, Ragweed, and Kochia 100% correctly. The lowest prediction percentage was obtained for Redroot Pigweed (90.9%), out of which 9.09% of images were predicted incorrectly as Sugar beet. However, the S1modelS2tedata confusion matrix did not have classes with 100% prediction accuracy, where Horseweed had the highest prediction accuracy of 97.9% and Palmer Amaranth had the lowest prediction accuracy of 28.3%. Moreover, Redroot Pigweed (40%) also had the poorest performance after Palmer Amaranth. For both Palmer Amaranth and Redroot Pigweed, a large percentage of images were wrongly predicted as Sugar beets causing lower prediction accuracy (Fig. 8b). Besides these species, the remaining species had a prediction accuracy in between 79% and 98%.

Table 4
The VGG16-C model performance in terms of precision, recall, and f1-score metrics for S1tedata (**CmodelS1tedata**) and S2tedata (**CmodelS2tedata**).

	precision		recall		f1-score		number of test data	
	CmodelS1tedata	CmodelS2tedata	CmodelS1tedata	CmodelS2tedata	CmodelS1tedata	CmodelS2tedata	CmodelS1tedata	CmodelS2tedata
Horseweed	1.00	0.98	1.00	0.94	1.00	0.96	45	48
Palmer Amaranth	0.98	0.97	1.00	0.97	0.99	0.97	42	60
Redroot Pigweed	1.00	0.97	1.00	1.00	1.00	0.98	22	30
Ragweed	1.00	1.00	1.00	1.00	1.00	1.00	47	65
Waterhemp	0.98	0.98	1.00	1.00	0.99	0.99	45	62
Canola	0.99	0.99	1.00	1.00	1.00	0.99	175	214
Kochia	1.00	0.95	1.00	0.97	1.00	0.96	31	38
Sugar beets	1.00	0.99	0.98	0.97	0.99	0.98	170	184
macro avg	**0.99**	**0.98**	**1.00**	**0.98**	**1.00**	**0.98**	577	**701**
weighted avg	**0.99**	**0.98**	**0.99**	**0.98**	**0.99**	**0.98**	577	**701**

Table 5
The ResNet50-S1 model performance in terms of precision, recall, and f1-score metrics for S1tedata (**S1modelS1tedata**) and S2tedata (**S1modelS2tedata**).

	precision		recall		f1-score		number of test data	
	S1modelS1tedata	S1modelS2tedata	S1modelS1tedata	S1modelS2tedata	S1modelS1tedata	S1modelS2tedata	S1modelS1tedata	S1modelS2tedata
Horseweed	1.00	0.85	1.00	0.98	1.00	0.91	45	48
Palmer Amaranth	0.95	0.94	1.00	0.28	0.98	0.44	42	60
Redroot Pigweed	0.95	0.63	0.91	0.40	0.93	0.49	22	30
Ragweed	1.00	1.00	1.00	0.92	1.00	0.96	47	65
Waterhemp	1.00	0.72	0.91	0.79	0.95	0.75	45	62
Canola	0.98	0.96	0.99	0.83	0.99	0.89	175	214
Kochia	1.00	0.72	1.00	0.82	1.00	0.77	31	38
Sugar beet	0.98	0.68	0.99	0.94	0.98	0.79	170	184
macro avg	**0.98**	**0.81**	**0.97**	**0.75**	**0.98**	**0.75**	577	**701**
weighted avg	**0.98**	**0.84**	**0.98**	**0.81**	**0.98**	**0.80**	577	**701**

The ResNet50-S2 model performance for both S1tedata and S2tedata in terms of f1-score is depicted in Table 6. When the ResNet50-S2 model was evaluated on S2tedata, the macro average f1-score and weighted f1-score was found to be 95% and 96%, respectively. However, when the model was evaluated on S1tedata, performance declined significantly with a macro and weighted f1-score value of 68% and 63%. There could be several reasons for this, but the major reason behind this was due to completely different backgrounds on test images than that of training images. When evaluating the ResNet50-S2 model with S2tedata, the highest f1-score value was obtained for Horseweed (100%) and the lowest f1-score value was obtained for Redroot Pigweed (87%). The severe decline in performance of the ResNet50-S2 model on S1tedata was due to the significant decline of f1-score value in Sugar beet, Palmer Amaranth, and Redroot Pigweed below 45% (Table 6). For Sugar beet, the decline in the f1-score value is due to the decline of the recall value to 21%, whereas for Redroot Pigweed and Palmer Amaranth, both the precision and recall values declined below 51%.

The confusion matrix of the ResNet50-S2 model for S2tedata and S1tedata is shown in Fig. 8 c and d, respectively. For S2modelS2tedata, Horseweed, Waterhemp, and Canola test images were predicted 100% correctly, where all the classes had greater than 90% prediction accuracy, except Palmer Amaranth (80%). For Palmer Amaranth, 6.67% of the images were misclassified as Sugar beet and 5% of the images were misclassified as Redroot Pigweed. When the ResNet50-S2 model was evaluated on the S1tedata, the model performance declined compared to the S2tedata. In the confusion matrix for S1tedata in Fig. 8d, Sugar beet prediction accuracy was significantly lower due to misclassification of Sugar beet images as Canola. Similarly, the prediction accuracy for Palmer Amaranth and Redroot Pigweed was below 51%. However, Horseweed, Canola, and Ragweed had an outstanding performance with a prediction accuracy over 97%. The overall pattern of performance degradation on ResNet50-S2 model was similar to that of ResNet50-S1 model when the model was tested on different background test data.

The ResNet50-C model was evaluated on both S1tedata and S2tedata to see the performance difference in CmodelS1tedata and CmodelS2tedata. Table 7 shows the precision, recall, and f1-score values for ResNet50-C model when tested with both S1tedata and S2tedata. Macro average f1-score for both S1tedata and S2tedata was obtained as 93% and weighted average f1-score was obtained as 95% for both test data. Therefore, there was no significant performance difference when using two test data because the model was trained using both background scenario training and validation data. For CmodelS1tedata, f1-score values for eight classes range in between 79% (Palmer Amaranth) and 100% (Horseweed, Ragweed, and Kochia). Similarly, in CmodelS2tedata, f1-score value was in between 79% (Palmer Amaranth) and 100% (Horseweed). Palmer Amaranth had the lowest f1-score due to having recall value below 70% for CmodelS1tedata and CmodelS2tedata.

Furthermore, Fig. 8e and f shows the confusion matrix with percentage of the images correctly classified and misclassified for ResNet50-C model. For ResNet50 CmodelS1tedata, 100% of the Horseweed, Ragweed, and Kochia images were classified correctly. Palmer Amaranth had the poorest prediction accuracy of 69% and most of Palmer Amaranth (14.25%) images were misclassified as the Waterhemp. In addition to Palmer Amaranth, Redroot Pigweed also had poor prediction accuracy of 72.7% because 22.7% of the images were misclassified as Sugar beet. However, for ResNet50 CmodelS2tedata, only Palmer Amaranth had poor performance with prediction accuracy of 65% (Fig. 8f). Remaining classes had prediction accuracy of over 96% percent except Redroot Pigweed with prediction accuracy of 86.7%. This shows overall prediction accuracy was impressive when both background data scenarios were mixed.

3.4. Model's comparisons and statistical significance under two background scenarios

Model performance statistical comparison based on average f1-scores and average accuracies was performed. The average accuracy of the six testing scenarios for both VGG16 and ResNet50 is depicted in Fig. 9. The graph exhibits the performance degradation of S1model and S2model in both deep learning CNN architectures when the model was tested on test data with different background from the data used in building a model. The graph in Fig. 9 shows how each classes' accuracy deviated from the average accuracy for both CNN architectures. There was higher deviation in the accuracy in the classes from the mean for both S1 model and S2 model when the model was tested on test data with a different background from the model building data. Conversely, S1 model and S2 model tested on test data with a similar background as the model building data performed well, and the deviation from mean accuracy was quite small.

The percentage of model performance degradation is depicted in Table 8 with its statistical significance. In the VGG16 model, mean accuracy declined by 18.97% for S1 model and 17.88% for S2 model when models were tested on test data with different background scenarios from the model building data. The paired-*t*-test results for both VGG16-S1 model ($t(7) = 2.3$, $p = 0.027$), and VGG16-S2 model ($t(7) = 2.24$, $p = 0.03$) showed that the differences in mean accuracies were significant at an alpha level of 0.05. The accuracy model results showed that VGG16-S2 model was able to generalize well on completely different sets of test data. However, accuracy may not give an exact picture of the model performance since model performance was compared based on f1-score to make a conclusion. The ResNet50 model also shows similarities with VGG16 model. The average accuracy of the ResNet50-S1 model and ResNet50-S2 model declined by 23.52% and 25.56% for S2tedata and S1tedata, respectively. These values suggest a greater decline of the model performance in ResNet50 than the VGG16.

The average accuracy explains the percentage of the correct classification of images but does not depict the level of wrong classification of images into certain classes. Hence, the average f1-score was used as the main performance metric. The average f1-score error bar plot in Fig. 10

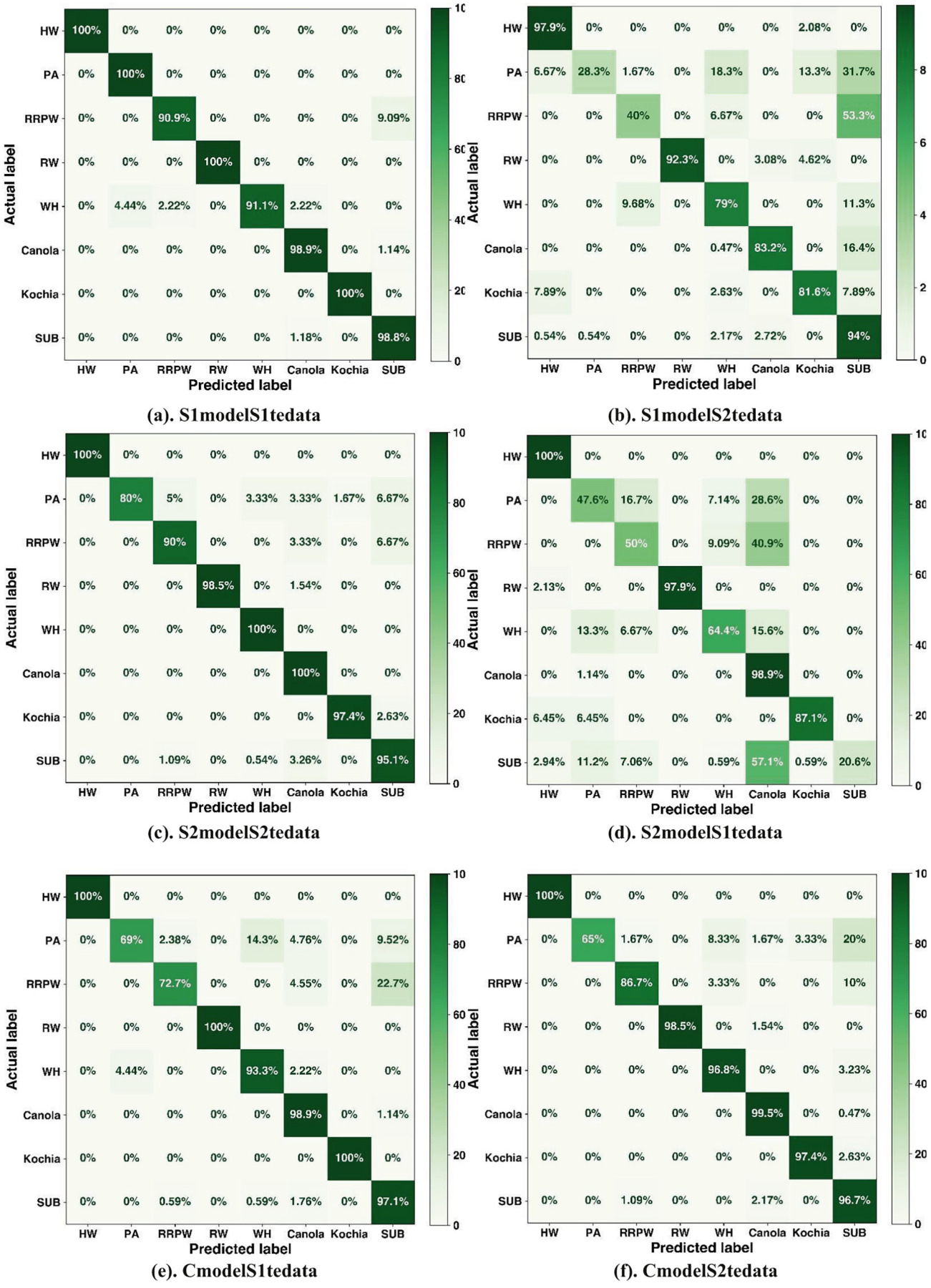


Fig. 8. The ResNet50 model test results in terms of confusion matrix. (a). S1model tested with S1tedata (b). S1model tested with S2tedata (c). S2model tested with S2tedata (d). S2model tested with S1tedata (e). Cmodel tested with S1tedata (f). Cmodel tested with S2tedata (HW: Horseweed, PA: Palmer Amaranth, RRPW: Redroot Pigweed, RW: Ragweed, WH: Waterhemp, SUB: Sugar beet).

Table 6
The ResNet50-S2 model performance in terms of precision, recall, and f1-score metrics for S1tedata (**S2modelS1tedata**) and S2tedata (**S2modelS2tedata**).

	precision		recall		f1-score		number of test data	
	S2modelS2tedata	S2modelS1tedata	S2modelS2tedata	S2modelS1tedata	S2modelS2tedata	S2modelS1tedata	S2modelS2tedata	S2modelS1tedata
Horseweed	1.00	0.85	1.00	1.00	1.00	0.92	48	45
Palmer Amaranth	1.00	0.41	0.80	0.48	0.89	0.44	60	42
Red Root Pigweed	0.84	0.33	0.90	0.50	0.87	0.40	30	22
Ragweed	1.00	1.00	0.98	0.98	0.99	0.99	65	47
Waterhemp	0.95	0.83	1.00	0.64	0.98	0.73	62	45
Canola	0.96	0.58	1.00	0.99	0.98	0.73	214	175
Kochia	0.97	0.96	0.97	0.87	0.97	0.92	38	31
Sugar beet	0.96	1.00	0.95	0.21	0.96	0.34	184	170
macro avg	**0.96**	**0.81**	**0.95**	**0.81**	**0.95**	**0.68**	**701**	577
weighted avg	**0.96**	**0.83**	**0.96**	**0.78**	**0.96**	**0.63**	**701**	577

Table 7
The ResNet50-C model performance in terms of precision, recall, and f1-score metrics for S1tedata (**CmodelS1tedata**) and S2tedata (**CmodelS2tedata**).

	precision		recall		f1-score		number of test data	
	CmodelS1tedata	CmodelS2tedata	CmodelS1tedata	CmodelS2tedata	CmodelS1tedata	CmodelS2tedata	CmodelS1tedata	CmodelS2tedata
Horseweed	1.00	1.00	1.00	1.00	1.00	1.00	45	48
Palmer Amaranth	0.94	1.00	0.69	0.65	0.79	0.79	42	60
Redroot Pigweed	0.89	0.90	0.73	0.87	0.80	0.88	22	30
Ragweed	1.00	1.00	1.00	0.98	1.00	0.99	47	65
Waterhemp	0.86	0.91	0.93	0.97	0.89	0.94	45	62
Canola	0.96	0.97	0.99	1.00	0.97	0.98	175	214
Kochia	1.00	0.95	1.00	0.97	1.00	0.96	31	38
Sugar beet	0.94	0.90	0.97	0.97	0.95	0.93	170	184
macro avg	**0.95**	**0.95**	**0.91**	**0.93**	**0.93**	**0.93**	**577**	**701**
weighted avg	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**577**	**701**

depicts the pattern of decline in the performance for both VGG16 and ResNet50 models when the models were tested on test data with different backgrounds than that of the model building data. Besides this, there was higher deviation of the f1-score values of classes from the mean for both S1model and S2model when the model was tested on test data with a different background from the model building data. Fig. 10 also shows a lower performance of the ResNet50-S1 model and ResNet50-S2 model than that of the VGG16.

Table 8 shows the average percentage of decline in f1-scores when modelsVGG16-S1, VGG16-S2, ResNet50-S1, and ResNet50-S2 were tested on test data with a different background from the model building data. The average f1-score of VGG16-S1 model and VGG16-S2 model declined by 16.83% and 21.22%, respectively, whereas the average f1-score of the ResNet50-S1 model and ResNet50-S2 model declined by 23.37% and 28.40%, respectively. This means S1model was more generalized than that of S2model in both VGG16 and ResNet50 because of its higher performance on unseen test data with a different background. Moreover, the paired-*t*-test results shows that the differences in f1-scores between models VGG16-S1 ($t(7) = 2.95$, $p = 0.01$), VGG16-S2 ($t(7) = 3.99$, $p = 0.0026$), ResNet50-S1 ($t(7) = 3.689$, $p = 0.0039$), and ResNet50-S2 ($t(7) = 3.42$, $p = 0.0056$) were significant using an alpha level of 0.05 (Table 8).

The decline in model performance between VGG16-S1, VGG16-S2, ResNet50-S1, and ResNet50-S2 was addressed by the training of the model using both background scenarios training data to develop the VGG16-C model and ResNet50-C model. The improvement of the model performance of Cmodel from S1model and S2model is depicted for both VGG16 and ResNet50 models in Fig. 9 and Fig. 10. However,

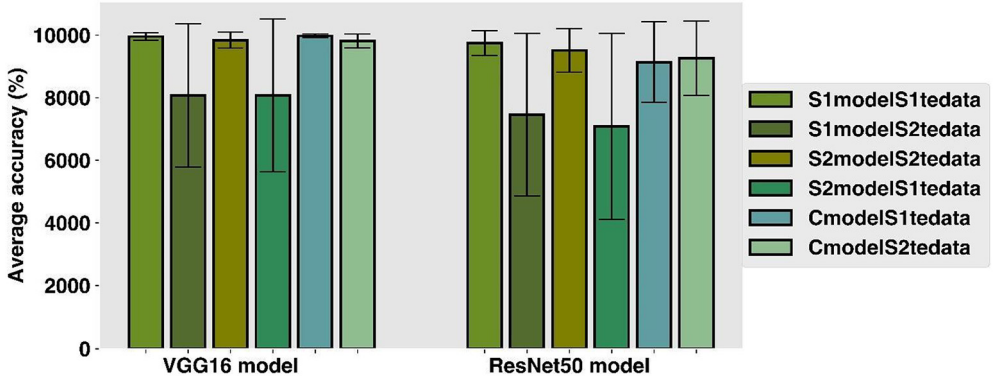


Fig. 9. The VGG16 and ResNet50 average accuracy error bar graph plot for six types of testing: S1modelS1tedata: S1model tested on S1tedata, S1modelS2tedata; S1model tested on S2tedata, S2modelS2tedata; S2model tested on S2tedata S2modelS1tedata; S2model tested on S1tedata, CmodelS1tedata; Cmodel tested on S1tedata, and CmodelS2tedata; Cmodel tested on S2tedata.

Table 8
Model performance degradation when model was tested on completely different background scenario data; n is the number of classes, SD is the standard deviation, PE is percentage of error when S1model and S2model model were tested with different background test data, and *p*-value gives the significance of the model performance degradation using paired-t-test.

	n	Mean (%)	SD (%)	n	Mean (%)	SD (%)	PE (%)	p-value
VGG16								
	S1modelS1tedata			S1modelS2tedata				
F1-score	8	99.50	0.71	8	82.75	15.86	16.83	0.011
Accuracy	8	99.56	1.16	8	80.66	22.88	18.97	0.027
	S2modelS2tedata			S2modelS1tedata				
F1-score Accuracy	8	98.38	1.69	8	77.5	15.31	21.22	0.003
	8	98.36	2.53	8	80.78	24.37	17.88	0.030
ResNet50								
	S1modelS1tedata			S1modelS2tedata				
F1-score Accuracy	8	97.88	2.59	8	75	19.09	23.37	0.004
	8	97.46	4.02	8	74.54	25.95	23.52	0.018
	S2modelS2tedata			S2modelS1tedata				
F1-score Accuracy	8	95.50	4.81	8	68.38	25.85	28.40	0.006
	8	95.13	6.10	8	70.81	29.70	25.56	0.017

Table 9
Model performance improvement when model was tested on combined data from both background scenarios; n is the number of classes, SD is the standard deviation, PI is the percentage of improvement when model was developed with combined dataset of two background scenario, and *p*-value gives the significance of the model performance improvement using paired-*t*-test.

	n	Mean (%)	SD (%)	n	Mean (%)	SD (%)	PI (%)
VGG16							
	S1modelS2tedata			CmodelS2tedata			
F1-score	8	82.75	15.86	8	97.88	1.46	18.28
Accuracy	8	80.66	22.88	8	98.15	2.27	21.66
	S2modelS1tedata			CmodelS1tedata			
F1-score Accuracy	8	77.5	15.31	8	99.63	0.52	28.55
	8	80.78	24.37	8	99.78	0.64	23.52
ResNet50							
	S1modelS2tedata			CmodelS2tedata			
F1-score Accuracy	8	75.00	19.09	8	93.38	6.97	24.5
	8	74.54	25.95	8	92.58	11.90	24.20
	S2modelS1tedata			CmodelS1tedata			
F1-score Accuracy	8	68.38	25.85	8	92.5	8.83	35.28
	8	70.81	29.70	8	91.375	12.90	29.04

the deviation of the accuracy and f1-score from the mean was higher for the ResNet50 than the VGG16 model. Moreover, the performance improvement is clearly depicted on Table 9 for both VGG16 and ResNet50 in terms of accuracy and f1-score. The improvement of the model performance of the VGG16-C model on S2tedata were 18.28% in terms of f1-score and 21.66% in terms of accuracy from the VGG16-S1 model. Whereas the improvement of the model performance of the VGG16-C model on S1tedata were 28.55% in terms of f1-score and 23.52% in terms of accuracy from the VGG-S2 model. The performance improvement pattern in the ResNet50 model was also same as that of VGG16. The improvement of the model performance of the ResNet50-C model on S2tedata were 24.5% in terms of f1-score and 24.20% in terms of accuracy from the ResNet50-S1 model. Whereas the improvement of the model performance of the ResNet50-C model on S1tedata were 35.28% in terms of f1-score and 29.04% in terms of accuracy from the ResNet50-S2 model. The performance improvement was higher for S2 model than S1model for both VGG16 and ResNet50, which was expected as the S2 model performance was lower than the S1 model in both VGG16 and ResNet50. The model developed from non-uniform (S1) image background dataset performed better than uniform (S2) image background dataset. When the model from combined image background dataset was tested on uniform and non-uniform test datasets, the weed identification performance improved better at non-uniform image background test dataset.

4. Discussion

In this research article, an applications of deep learning architecture to study the performance of the model with different image background conditions under greenhouse condition was performed on weed and crop classification task. Same crop and weed plants were planted in pots and two different image backgrounds scenarios as non-uniform background scenario (S1) and uniform background scenario (S2) was created. The Deep Learning (DL) models that were developed from the two scenarios was compared in terms of developing a more robust model and to understand how the models perform on different image analysis conditions than the condition the model was created. The DL robustness was checked for completely unseen images with different image background than that of images used for model training. The results confirmed that model performance declined in both types of background scenarios when model was cross-tested on images with different background scenario than that of images used during model training. In weed classification study, Espejo-Garcia et al. (2020) performed plant segmentation to check if deep learning model after

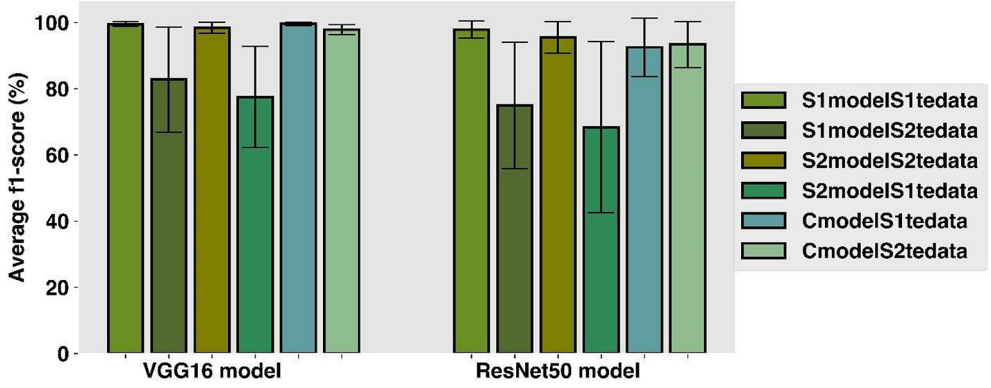


Fig. 10. The VGG16 and ResNet50 accuracy f1-score error bar graph plot for six types of testing. S1modelS1tedata: S1model tested on S1tedata, S1modelS2tedata; S1model tested on S2tedata, S2modelS2tedata; S2model tested on S2tedata S2modelS1tedata; S2model tested on S1tedata, CmodelS1tedata; Cmodel tested on S1tedata, and CmodelS2tedata; Cmodel tested on S2tedata.

background removal performs better and found that result did not improve. The study performed in this article is different from the study performed by Espejo-Garcia et al. (2020), where different background condition was created by the application of the black pebble on the top of the promix mixture. In this study, two different background images was created by application of black gravels in same crop and weed plants which is different from Espejo-Garcia et al. (2020) study. Beside the background in image, black gravel also affects the lighting around the plant due to more absorption of light by black surfaces than other color surfaces (Stuart-Fox et al., 2017). The study's results performed for plant disease detection by Ferentinos (2018) was in correspondence to our study results in weed classification. In their research, model performance decline from 99% to 68% when model trained with field condition image was tested with laboratory-conditions images. The performance decline even more to 33% when types of images for model training and testing was reversed.

The VGG16S1 model (model trained on non-uniform (S1) data) in terms of f1-score declined by 16.83% when tested with S2tedata (uniform (S2) test data), whereas the VGG16 S2model (model trained on uniform (S2) data) performance declined by 21%, when the model was tested with S1tedata (non-uniform (S1) test data). Similar pattern of declination was observed in ResNet50, where ResNet50-S1 model performance in terms of F1-score declined by 23.37% and ResNet50-S2 model performance declined by 28.40%. This decline in performance was due to testing the models on image sets with different backgrounds than that of images used in model training. This result also shows the cause of difficulty of the implementation of deep learning models in real time weed detection applications due to different soil background conditions such as color and texture. The model's performance was improved by 18% to 35% in both CNN architectures when both types of images were used to build a model. This suggests to used wider variety of training weed and training data collected from different geographical area, cultivation condition, imaging sensors, and image capturing mode as suggested by Ferentinos (2018) in plant disease detection study.

In addition, studies in research and reviews on the use of digital images on weed identification using DL have shown no use of six weed species (Horseweed, Palmer Amaranth, Redroot Pigweed, Kochia, Ragweed, and Waterhemp) and two crop species (Canola and Sugar beet) together to build a model (Espejo-Garcia et al., 2020; Hasan et al., 2021; Khan et al., 2021; Olsen et al., 2019). However, these studies have limitation focus on the performance of the DL model when images have different background conditions and did not involve popular weed and crop species that found popular in the Midwest United States Region. In our study, the average f1-score values from 92% to 99% was observed, when the model was developed with combined datasets which is in accordance with the research finding by Espejo-Garcia et al. (2020), Hu et al. (2020), Khan et al. (2021), Le et al. (2020), Olsen et al. (2019), and Razfar et al. (2022) in weed detections. In comparison, our research was performed for six species of weeds that are prevalent on North Dakota and two important crop species grown on North Dakota. Study on use of machine learning and deep learning in North Dakota based crop and weed species classification was performed in our previous research article (GC et al., 2022), which was mainly focused on weed classification under different crop production system and did not involve any study about impact of image background on model performance. The result with combined datasets seems outstanding despite three weed species (Waterhemp, Palmer Amaranth, and Redroot Pigweed) with similar characteristics being used for building a model.

5. Conclusion

This research investigated the effects of image background in Deep Learning (DL) models, specifically compared Convolutional Neural Networks (CNNs) of VGG16 and ResNet16 on non-uniform and uniform image background conditions. The analysis exhibited a decline in weed identification performance by 16% to 28% in the VGG16 and ResNet50 model, when models were tested on images with different background than that of model training. However, a model trained with combined datasets of two background scenarios, was able to achieve the f1-score value of 92% to 99% with eight classes of weed and crop images. The performance parameters for each weed and crop species in terms of accuracy, f1-score, and confusion matrix were greater than expected. The findings of this study suggest that DL models developed from uniform image background could not improve the performance of deep learning models to identify weeds in unseen environmental conditions. From this study we suggest that similar kind of background soil environment as that of field should be created inside the greenhouse to be able to apply the deep learning-based weed detection model in field. In addition, it seems including greater variances in image background for the models seems working better when tested in individual conditions (when there is single variable such as non-uniform, or uniform, or clay, or sandy soil, or volcanic ash) than just developing the model in same condition and testing in same condition. Future work will focus on fusion of greenhouse and field images to create a field environment background condition to build robust weed detection deep learning model with localization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This material is based upon work partially supported by the USDA - Agricultural Research Service, agreement number 58-6064-8-023. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture. This work is/was supported by the USDA-National Institute of Food and Agriculture, Hatch project number ND01487. In addition, we want to thank NDSU Plant Science department's research specialists and Agricultural and Biosystems Engineering department graduate students (Arjun Upadhyay and Billy Ram) for helping with some of the greenhouse sample processes in this study.

References

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

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data 81 (8), 1–74. https://doi.org/10.1186/S40537-021-00444-8.

Arya, D., Maeda, H., Ghosh, S.K., Toshniwal, D., Mraz, A., Kashiyama, T., Sekimoto, Y., 2020. Transfer Learning-Based Road Damage Detection for Multiple Countries.

Bradski, G., 2000. The OpenCV Library. Dr. Dobb's J. Softw. Tools.

Bryson, C.T., DeFelice, M.S., 2010. Weeds of the Midwestern United States and Central Canada, A Wormsloe Foundation Nature Book. University of Georgia Press, Athens.

Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Fountas, S., Vasilakoglou, I., 2020. Towards weeds identification assistance through transfer learning. Comput. Electron. Agric. 171, 105306. https://doi.org/10.1016/J.COMPAG.2020.105306.

Ferentinos, K.P., 2018. Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. 145, 311–318. https://doi.org/10.1016/J.COMPAG.2018.01.009.

GC, S., Saidul Md, B., Zhang, Y., Reed, D., Ahsan, M., Berg, E., Sun, X., 2021. Using deep learning neural network in artificial intelligence technology to classify beef cuts. Front. Sensors 0, 5. https://doi.org/10.3389/FSENS.2021.654357.

GC, S., Zhang, Y., Koparan, C., Ahmed, M.R., Howatt, K., Sun, X., 2022. Weed and crop species classification using computer vision and deep learning technologies in greenhouse conditions. J. Agric. Food Res. 9, 100325. https://doi.org/10.1016/J.JAFR.2022.100325.

Hasan, A.S.M.M., Sohel, F., Diepeveen, D., Laga, H., Jones, M.G.K., 2021. A survey of deep learning techniques for weed detection from images. Comput. Electron. Agric. 184, 106067. https://doi.org/10.1016/J.COMPAG.2021.106067.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December, pp. 770–778 https://doi.org/10.1109/CVPR.2016.90.

Holzner, W., 1982. Concepts, categories and characteristics of weeds. Biology and Ecology of Weeds. Springer, Netherlands, pp. 3–20 https://doi.org/10.1007/978-94-017-0916-3_1.

Hu, K., Coleman, G., Zeng, S., Wang, Z., Walsh, M., 2020. Graph weeds net: a graph-based deep learning method for weed recognition. Comput. Electron. Agric. 174, 105520. https://doi.org/10.1016/J.COMPAG.2020.105520.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 1–54. https://doi.org/10.1186/S40537-019-0192-5/TABLES/18.

Khan, S., Tufail, M., Khan, M.T., Khan, Z.A., Anwar, S., 2021. Deep learning-based identification system of weeds and crops in strawberry and pea fields for a precision agriculture sprayer. Precis. Agric. 22, 1711–1727. https://doi.org/10.1007/S11119-021-09808-9/TABLES/3.

Kingma, D.P., Ba, J., 2015. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6.

Kulawardhana, R.W., 2011. Remote sensing of vegetation: principles, techniques and applications. By Hamlyn G. Jones and Robin a Vaughan. J. Veg. Sci. 22, 1151–1153. https://doi.org/10.1111/J.1654-1103.2011.01319.X.

Le, V.N.T., Ahderom, S., Alameh, K., 2020. Performances of the LBP based algorithm over CNN models for detecting crops and weeds with similar morphologies. Sensors (Basel). 20. https://doi.org/10.3390/S20082193.

Li, Y., Guo, Z., Shuang, F., Zhang, M., Li, X., 2022. Key technologies of machine vision for weeding robots: a review and benchmark. Comput. Electron. Agric. 196, 106880. https://doi.org/10.1016/J.COMPAG.2022.106880.

Ma, X., Wu, H., Jiang, W., Ma, Yajie, Ma, Yan, 2015. Interference between redroot pigweed (*Amaranthus retroflexus* L.) and Cotton (*Gossypium hirsutum* L.): growth analysis. PLoS One 10. https://doi.org/10.1371/JOURNAL.PONE.0130475.

Oerke, E., 2005. Crop losses to pests. J. Agric. Sci. 144, 31–43.

Oerke, E., 2005. Crop losses to pests. J. Agric. Sci. 144 (1), 31–43.

Olsen, A., Konovalov, D.A., Philippa, B., Ridd, P., Wood, J.C., Johns, J., Banks, W., Girgenti, B., Kenny, O., Whinney, J., Calvert, B., Azghadi, M.R., White, R.D., 2019. DeepWeeds: a multiclass weed species image dataset for deep learning. Sci. Report. 91 (9), 1–12. https://doi.org/10.1038/s41598-018-38343-3.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in {P} ython. J. Mach. Learn. Res. 12, 2825–2830.

Razfar, N., True, J., Bassiouny, R., Venkatesh, V., Kashef, R., 2022. Weed detection in soybean crops using custom lightweight deep learning models. J. Agric. Food Res. 8, 100308. https://doi.org/10.1016/J.JAFR.2022.100308.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.

Soltani, N., Dille, J.A., Burke, I.C., Everman, W.J., Vangessel, M.J., Davis, V.M., Sikkema, P.H., 2017. Perspectives on potential soybean yield losses from weeds in North America. Weed Technol. 31, 148–154. https://doi.org/10.1017/WET.2016.2.

Stuart-Fox, D., Newton, E., Clusella-Trullas, S., 2017. Thermal consequences of colour and near-infrared reflectance. Philos. Trans. R. Soc. B Biol. Sci. 372. https://doi.org/10.1098/RSTB.2016.0345.

Velumani, K., Lopez-Lozano, R., Madec, S., Guo, W., Gillet, J., Comar, A., Baret, F., 2021. Estimates of maize plant density from UAV RGB images using faster-RCNN detection model: impact of the spatial resolution. Plant Phenomics (Washington, D.C.). Doi: 10.34133/2021/9824843.

Wu, Z., Chen, Y., Zhao, B., Kang, X., Ding, Y., 2021. Review of weed detection methods based on computer vision. Sensors (Basel). 21. https://doi.org/10.3390/S21113647.