Article

Effect of Hyperparameter Tuning on the Performance of YOLOv8 for Multi Crop Classification on UAV Images

Oluibukun Gbenga Ajayi [1,2,*], **Pius Onoja Ibrahim [2]** and **Oluwadamilare Samuel Adegboyega [2]**

1   Department of Land and Spatial Sciences, Namibia University of Science and Technology,
    Windhoek Private Bag 13388, Namibia
2   Department of Surveying and Geoinformatics, Federal University of Technology Minna, Minna P.M.B. 65,
    Niger State, Nigeria; pius.onoja@futminna.edu.ng (P.O.I.); adegboyegaoluwadamilare@gmail.com (O.S.A.)
*   Correspondence: oajayi@nust.na or ogbajayi@gmail.com

Abstract: This study investigates the performance of YOLOv8, a Convolutional Neural Network (CNN) architecture, for multi-crop classification in a mixed farm with Unmanned Aerial Vehicle (UAV) imageries. Emphasizing hyperparameter optimization, specifically batch size, the study's primary objective is to refine the model's batch size for improved accuracy and efficiency in crop detection and classification. Using the Google Colaboratory platform, the YOLOv8 model was trained over various batch sizes (10, 20, 30, 40, 50, 60, 70, 80, and 90) to automatically identify the five different classes (sugarcane, banana trees, spinach, pepper, and weeds) present on the UAV images. The performance of the model was assessed using classification accuracy, precision, and recall with the aim of identifying the optimal batch size. The results indicate a substantial improvement in classifier performance from batch sizes of 10 up to 60, while significant dips and peaks were recorded at batch sizes 70 to 90. Based on the analysis of the obtained results, Batch size 60 emerged with the best overall performance for automatic crop detection and classification. Although the F1 score was moderate, the combination of high accuracy, precision, and recall makes it the most balanced option. However, Batch Size 80 also shows very high precision (98%) and balanced recall (84%), which is suitable if the primary focus is on achieving high precision. The findings demonstrate the robustness of YOLOv8 for automatic crop identification and classification in a mixed crop farm while highlighting the significant impact of tuning to the appropriate batch size on the model's overall performance.

Keywords: precision agriculture; smart farming; convolutional neural network; YOLOv8; hyperparameter; multi-crop classification; unmanned aerial vehicles

1. Introduction

Agriculture has been the platform from which human sustainability evolves over time, even with the spontaneous increase in population in recent time to over seven billion [1]. The activities of farming have changed, and are still changing, from the known traditional ways to a more sophisticated approach with the aid of evolving technology [2]. Planting one kind of crop on a single area of land is changing to a system of mixed cropping on the same area of land (mixed farming), which is gradually becoming a common practice [3]. Mapping out crops in mixed farming for effective management to meet the purpose of better harvest is a crucial component of precision agriculture. Therefore, the task of optimized multi-crop classification in mixed farms has been a subject of interest for many agricultural researchers worldwide [4]. This task involves the identification and classification of multiple crop types in a single agricultural field.

The classification of multiple crops in mixed farming systems is a challenging yet vital task which provides insights into the dynamics of crop combinations and their impact on yields and sustainability [5]. While early studies on multi-crop classification focused on

traditional rule-based methods and required enormous resources and time to execute [6], with advancements in remote sensing technologies, machine learning, and deep learning the approach to this task has evolved significantly [7]. Modern techniques, including the use of satellite imagery, drones, and deep learning algorithms, have revolutionized multi-crop classification because these methods offer higher accuracy and scalability [8].

The benefits of accurate crop classification are multifaceted. It enables optimized resource allocation, where farmers can efficiently distribute resources such as water, nutrients, and pesticides to specific crops based on their individual needs. This approach supports precision agriculture practices, allowing for site-specific management and data-driven decision-making [9]. Furthermore, accurate classification aids in yield prediction, helping farmers to estimate and plan for future harvests more accurately [10]. Nonetheless, crop classification comes with its own set of challenges. Crops evolve over time, and change in appearance as they progress through different growth stages. This can make classification difficult, especially if the same field is used to plant different crops in various seasons. Additionally, mixed planting nearby can lead to overlapping canopies, making it more challenging to identify and classify crops accurately [11]. A range of tools and technologies are available for crop classification in mixed farming. High-resolution satellite imagery can provide an overview of the entire farm, assisting in large-scale classification [12]. Drones equipped with cameras or sensors can capture detailed images of crops from a closer vantage point. Spectral imaging, such as hyperspectral or multispectral sensors, can capture the unique spectral signatures of different crops, aiding in their identification [2].

In recent years, deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have become useful models in precision agriculture [7]. They can distinguish between different crops and even assess their health and growth stage based on training data [2,12]. YOLOv8 is a popular CNN architecture released on 10 January 2023, and has shown promising results in object detection and classification [13]. It is a rapid, precise, and user-friendly model designed for tasks such as object detection, segmentation, classification, and pose estimation. However, the performance of YOLOv8, like other architectures, is highly dependent on hyperparameters, which means that the choice of hyperparameters can significantly affect its performance. The algorithm has several hyperparameters that can be tuned to improve its performance, including the learning rate, batch size, number of epochs, and number of anchors.

Several studies have investigated the effect of hyperparameters on the performance of deep learning models for object detection and classification, but only a few have focused on performance for multi-crop classification in a mixed farm context. For example, in [14] the authors investigated the effect of different hyperparameters on the performance of YOLOv3 for crop classification using remote sensing data. They found that increasing the batch size improved the accuracy of the algorithm. The authors also found that increasing the number of anchors improved the recall rate of the algorithm. Similarly, a study by [15] investigated the effect of hyperparameters (training epochs) on the performance of YOLOv5 for multi-crop classification in a mixed farm. The authors used remote sensing data to classify five crop types and tuned the number of training epochs from 100 to 1000. They found that increasing the number of epochs improved the accuracy of the algorithm until the algorithm became saturated at the 700th epoch, when the model performance began to decline. Several studies have investigated the impact of batch size on the performance of deep learning models. In [16], the authors found that increasing batch size can lead to better generalization performance, while [17] found that larger batch sizes can lead to lower generalization errors. However, other studies have shown that the optimal batch size varies depending on the dataset and network architecture [18–20]. Several studies have investigated the impact of batch size on the performance of YOLOv3 and YOLOv4. In a study by [21], the authors found that larger batch sizes can lead to better performance on the COCO dataset. Similarly, [22] found that increasing batch size can lead to improved accuracy and speed in object detection tasks. Hence, this study focuses on investigating the

effect of varying batch sizes on the performance of YOLOv8 for multi-crop classification in a mixed farm.

Study Area

The study area is mixed-crop farmland situated in the Lapan Gwari neighborhood of Minna, the capital city of Niger State in Nigeria. Covering approximately 2.8 hectares, the site is geographically located between (9°31′33″ N, 6°30′02″ E) and (9°31′37″ N, 6°30′05″ E) at an elevation of about 250 m above sea level [15]. The farmland predominantly features loamy soil. The crops grown on the farm include Banana (Musa spp.), Pepper (Capsicum spp.), Spinach (Spinacia oleracea), and Sugarcane (Saccharum officinarum) [23,24]. Figure 1 describes the study area, providing a context from the broader region of Nigeria to the specific site in Lapan Gwari.
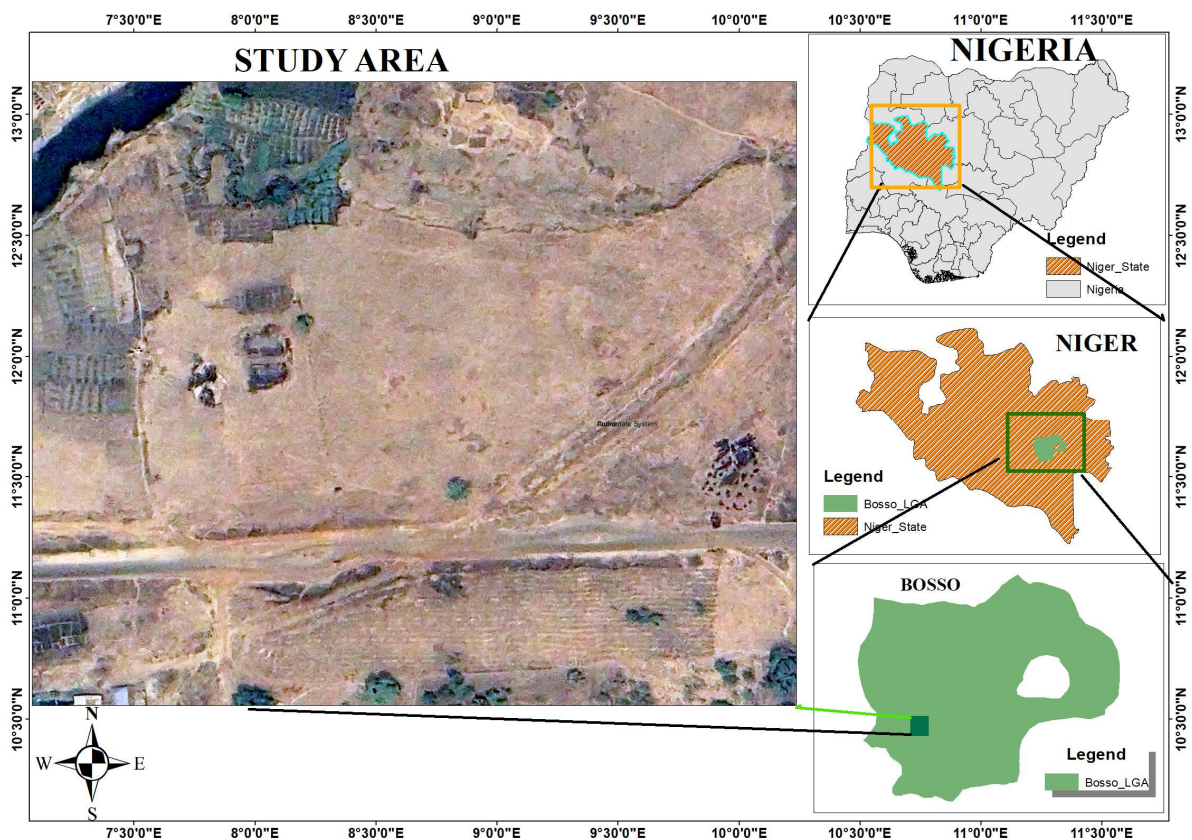

Figure 1. The study area in Lapan Gwari, Minna, Niger State, Nigeria [15,23].

2. Materials and Methods

The methodological approach adopted for the execution of this study is made up of four important steps: (i) data collection and preparation, (ii) preprocessing, (iii) model training, and (iv) evaluation metrics. The detailed breakdown of the approach used is outlined as follows:

i.   Acquisition of a diverse dataset of UAV images capturing various crop types found in the mixed farm and preparation of a labelled dataset from the acquired images.

ii.   Image resizing to standardized dimensions for compatibility with YOLOv8 input requirements. The initial 4000 × 3000-pixel images were resized to 416 × 416 pixels.

iii.   Implementation of the YOLOv8 architecture, known for its efficiency in object detection tasks, for crop classification and experimenting with different training batch sizes (10, 20, 30, 40, 50, 60, 70, 80, and 90) to investigate their impact on model performance.

iv.   Systematically analyzing results across various batch sizes to identify trends and variations in performance.

The procedure for developing and implementing a YOLOv8 architecture is described in Figure 2.
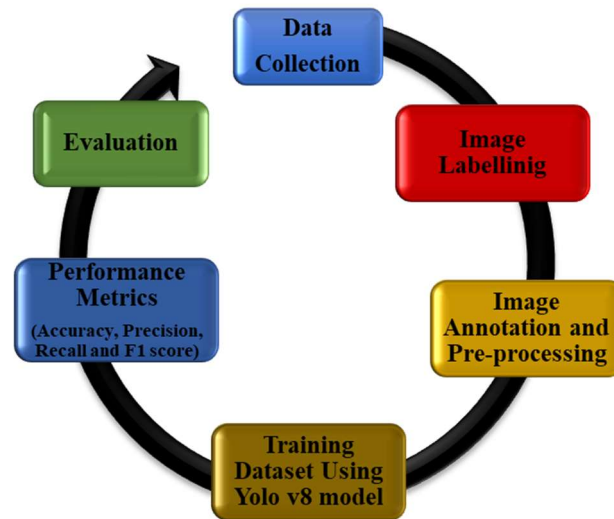
Figure 2. Procedure for developing and implementing YOLOv8 model.

2.1. Data Acquisition

The data used for this study was the same used by [15,23,24]. The data were acquired with the aid of a DJI Phantom 4 UAV (DJI, Shenzhen, China) equipped with an on-board RGB camera with 12 megapixels of camera resolution and a focal length of 5.74 mm. The drone was deployed at a flight altitude of 30 m and an average airspeed of 5 m/s with a front overlap of 75% and a side overlap of 65%. A total of 1488 images were collected for this study, out of which about 393 images were used, as these images were successfully annotated by the Computer Vision Annotation Tool (CVAT) (https://www.cvat.ai/), an interactive video and image annotation web-based tool for computer vision.

2.2. Image Preprocessing

The preprocessing procedures executed on the obtained images (Figure 3) are as discussed below.

Figure 3. Workflow for the data preprocessing.

Image resizing: Image resizing is a fundamental step aimed at standardization, which facilitates the storage of images in a NumPy array format suitable for input into a deep learning network. YOLOv8 typically requires input images to adhere to specific dimensions. In this case, the original images, initially sized at 4000×3000 pixels, were resized to 416×416 pixels. This process ensures uniformity in input dimensions, a prerequisite that enhances the ability of the YOLOv8 model to learn features effectively.

Normalization of Data: Normalizing pixel values is important for training deep learning models. It involves scaling pixel values to a standardized range, often between 0 and 1. This step helps to stabilize training while avoiding saturation of activation functions, improving

model robustness, and mitigating sensitivity to initial weights. It prevents issues related to convergence and ensures that input features are on a similar scale. This contributes to the stability, efficiency, and generalization ability of the YOLOv8 model during training, which promotes optimal performance on real-world data.

Data Augmentation: We artificially increased the diversity of the training dataset by applying a transformation function to the images. This helps the model to generalize better and improve its robustness. This diversification helps the YOLOv8 model to learn more invariant features, which makes it more resilient to variations in real-world scenarios. Essentially, data augmentation acts as a regularization technique that prevents overfitting and ensures that the model performs well on unseen data by exposing it to a broader range of variations during training.

Data Splitting: The dataset, consisting of 393 images, was divided into three subsets: 80% for training, 13% for validation, and 7% for testing (https://app.roboflow.com/final-year-project-avilz/image-annotation-vv5yx/1 (accessed on 21 September 2023)). The validation subset was used to evaluate the model's performance during the training phase, while the test subset was employed to measure the model's effectiveness after training. Figure 3 describes the workflow for the UAV image data preprocessing.

2.3. Implementation Architecture

The model design of YOLOv8's network architecture is presented in Figure 4. The model employs a feature extraction backbone network, typically the CSPDarknet53 architecture, to capture hierarchical features from UAV images. This feature extraction is important for identifying and classifying different crops, in this case including sugarcane, banana trees, spinach, pepper, and weeds [23,24]. To refine features obtained from the backbone, it incorporates a neck architecture such as Path Aggregation Network (PANet). This refinement process facilitates feature aggregation across different scales, allowing the model to effectively handle objects of varying sizes within the dataset. Anchor boxes, which are predefined bounding box shapes learned during training, enhance object localization and ensure accurate prediction of bounding box coordinates for each crop type. During the training process, YOLOv8 learns from labelled images and ground truth annotations, minimizing a loss function that considers both localization accuracy and classification performance [25,26]. After obtaining predictions, the model applies postprocessing techniques to filter out low-confidence detections and refine the final set of predictions for different crop types. This is essential for ensuring the accuracy of crop classification in precision agriculture [27].

2.4. Training Process

Before commencing model training, meticulous data preparation was carried out. The dataset, which included annotated images with labelled objects, was divided into three fundamental subsets: the training set, validation set, and test set, as described earlier. Each batch size variant underwent this same data partitioning process to ensure equitable training conditions. The YOLOv8 model's configuration files were tailored to accommodate the distinct requirements of each batch size. Each variant was configured to reference the appropriate dataset directories and batch size. The training process began with the initialization of the YOLOv8 model within the Google Colab environment. Importing the model and its dependencies ensured the availability of essential libraries and configurations. For the learning and identification tasks, a T4 GPU with 12 GB of RAM was used (NVIDIA GeForce GTX TITAN X). The study was conducted on a workstation operating on Ubuntu 18.04 with GPU acceleration using a virtual machine setup, while Python (Python 3.8) programming was employed for coding. The dataset consisted of images annotated with labelled bounding boxes to identify the crops and weeds. As the primary objective of this study was to examine the impact of batch size on the model's performance, a dedicated training script was executed for each batch size variant, referencing its specific dataset directories and configuration files. The training process entailed a set number of epochs,

during which the model adapted and learned from the data. The batch sizes used were 10, 20, 30, 40, 50, 60, 70, 80, and 90, allowing for comparative analysis. Comprehensive monitoring and metric analysis were conducted to assess the effect of batch size on the training process. Loss, an essential metric, was monitored continuously. The TensorBoard (https://www.tensorflow.org/tensorboard) tool facilitated real-time visualization of training metrics. The trained models for each batch size variant were evaluated using dedicated validation sets, which assessed their ability to generalize and detect objects accurately. Subsequently, the test set was employed to further validate the models' performance.

model	d(depth_multiple)	w(width_multiple)	r(ratio)
n	0.33	0.25	2.0
s	0.33	0.50	2.0
m	0.67	0.75	1.5
l	1.00	1.00	1.0
x	1.00	1.25	1.0
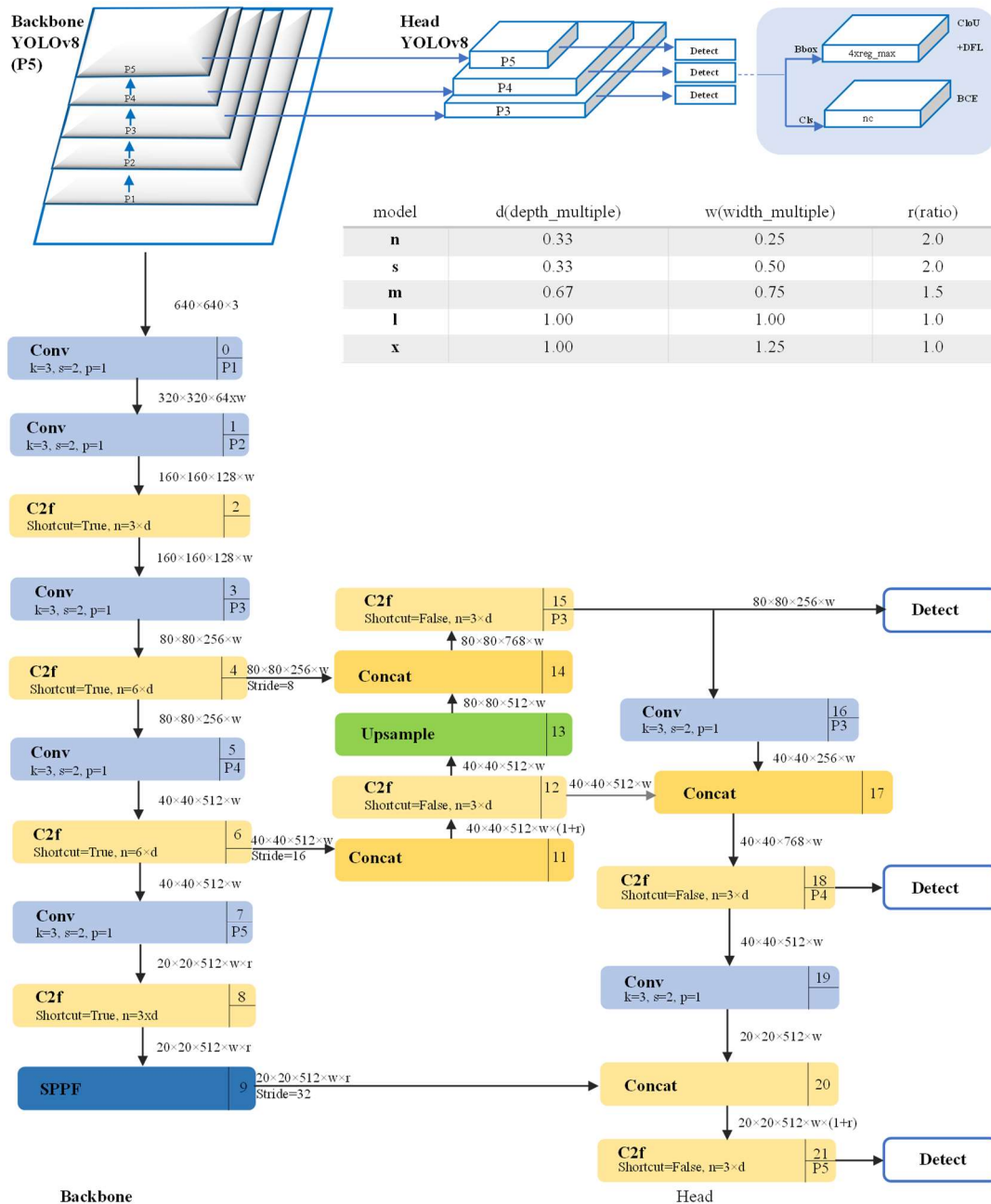
Figure 4. The design of YOLOv8's network architecture (adapted from https://github.com/ultralytics/ultralytics/issues/189 (accessed on 19 April 2024)).

2.5. Performance Evaluation

Several accuracy metrics are available to evaluate the performance of a deep learning model. In this study, the performance of the YOLO v8 algorithm was evaluated using

both the testing and validation datasets. The metrics used for this assessment included recall (R), accuracy (A), F1 score (F1), and precision (P). According to [28,29], these metrics are commonly employed in deep learning applications. The metrics used to assess the performance of the YOLOv8 model are discussed as follows:

i.   Accuracy (ACC): Accuracy is a fundamental metric that quantifies the model's overall correctness in its predictions. It is defined as the ratio of correctly classified objects to the total number of objects.

$$\text{ACC} = \frac{\textbf{\textit{True Positives}} + \textbf{\textit{True Negatives}}}{\textbf{\textit{Total Objects}}} \tag{1}$$

ii.   Precision (PR): Precision gauges the model's ability to make correct positive predictions. It is calculated as the ratio of true positive predictions to the total number of positive predictions.

$$\text{PR} = \frac{\textbf{\textit{True Positives}}}{\textbf{\textit{True Positives}} + \textbf{\textit{False Positives}}} \tag{2}$$

iii.   Recall (RC): Recall, also known as sensitivity or true positive rate, measures the model's capability to identify all relevant instances. It is defined as the ratio of true positive predictions to the total number of actual positive instances.

$$\text{RC} = \frac{\textbf{\textit{True Positives}}}{\textbf{\textit{True Positives}} + \textbf{\textit{False Negatives}}} \tag{3}$$

iv.   F1 Score (F1): The F1 score balances precision and recall to provide a single metric that quantifies the model's accuracy in detecting and classifying positive instances.

$$\text{F1} = 2 \cdot \frac{\textit{Precision. Recall}}{\textit{Precision} + \textit{Recall}} \tag{4}$$

v.   Precision Average (PR-AVG): The precision average is calculated as the arithmetic mean of precision values for each class or category in the classification problem:

$$\text{Precision Average (PR-AVG)} = \frac{1}{N}\sum_{i=1}^{N} Precision_i \tag{5}$$

where *N* represents the number of classes.

vi.   Recall Average (RC-AVG): Similarly, the recall average is computed as the average of recall values for each class.

$$\text{Recall Average (RC-AVG)} = \frac{1}{N}\sum_{i=1}^{N} Recall_i \tag{6}$$

For this study, precision, recall, and mean average precision (mAP) were employed to evaluate the performance of the model under different batch size conditions. mAP is derived from the precision–recall curve.

3. Results

The precision confidence curves for various batch sizes were analyzed in order to evaluate the performance of the network training process. The network was trained with batch sizes of 10, 20, 30, 40, 50, 60, 70, 80, and 90. The results indicated that the highest precision of 0.984 was achieved at a batch size of 80. This was followed by batch sizes of 30 with a precision of 0.964, 10 with 0.950, 70 with 0.937, 60 with 0.897, 40 with 0.870, 20 with 0.835, and finally batch size 50 with the lowest precision confidence of 0.821. Figure 5 shows the precision confidence training and validation output across the experimented batch sizes.
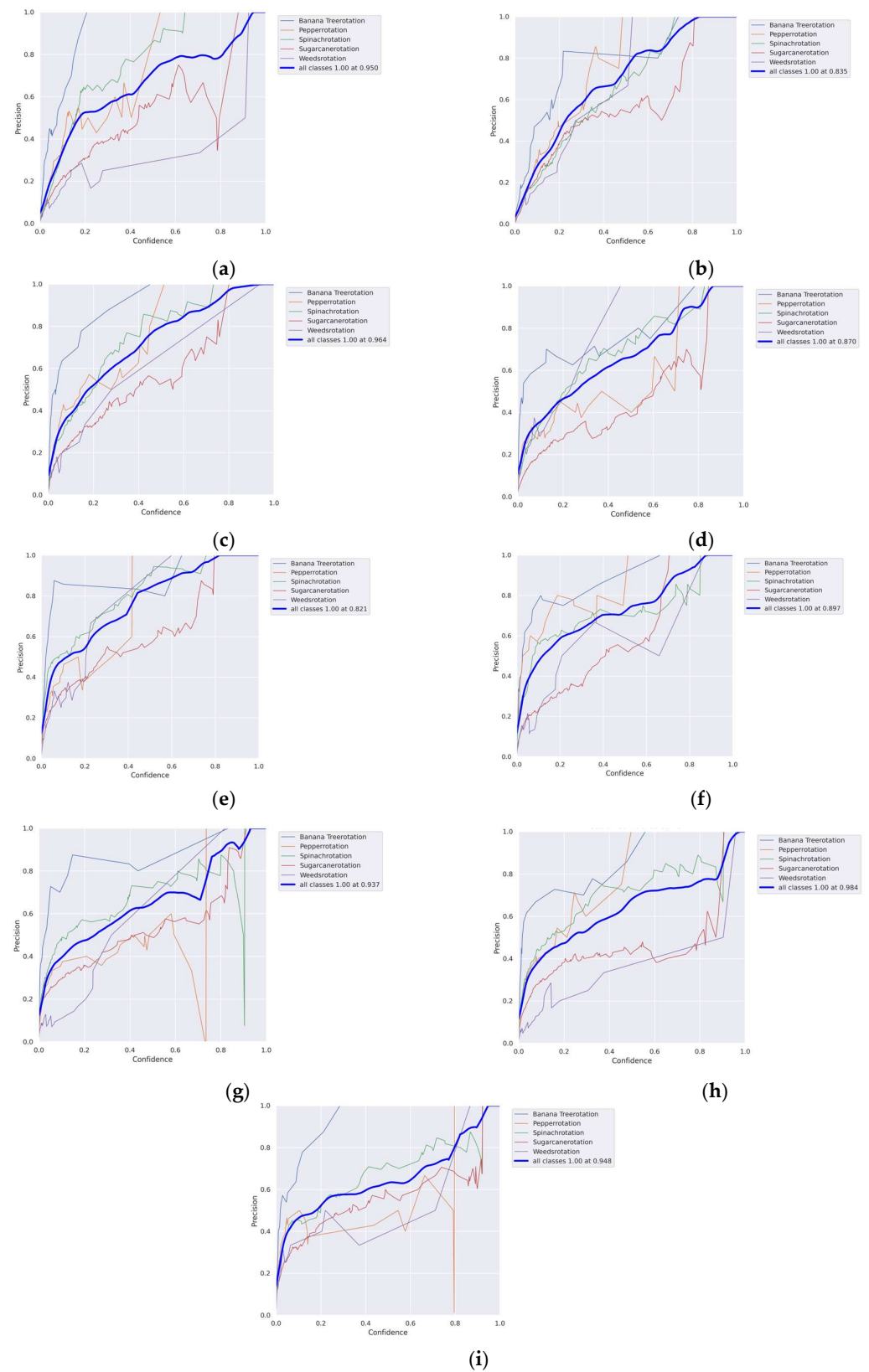
Figure 5. Precision confidence of the training and validation output across (**a**) batch size 10, (**b**) batch size 20, (**c**) batch size 30, (**d**) batch size 40, (**e**) batch size 50, (**f**) batch size 60, (**g**) batch size 70, (**h**) batch size 80, and (**i**) batch size 90.

3.1. Recall Confidence

It is essential to understand that setting the appropriate confidence threshold depends on the specific application and the trade-off between missing detections (false negatives) and accepting false positives. Figure 6 shows how the recall confidence level decreases as the batch sizes are increased. The obtained recall confidence values for each of the batch sizes were 0.840, 0.820, 0.780, 0.790, 0.750, 0.830, 0.770, 0.840, and 0.77 for batch sizes 10, 20, 30, 40, 50, 60, 70, 80, and 90, respectively. The highest recall confidence was recorded at batch size 10 and 80.
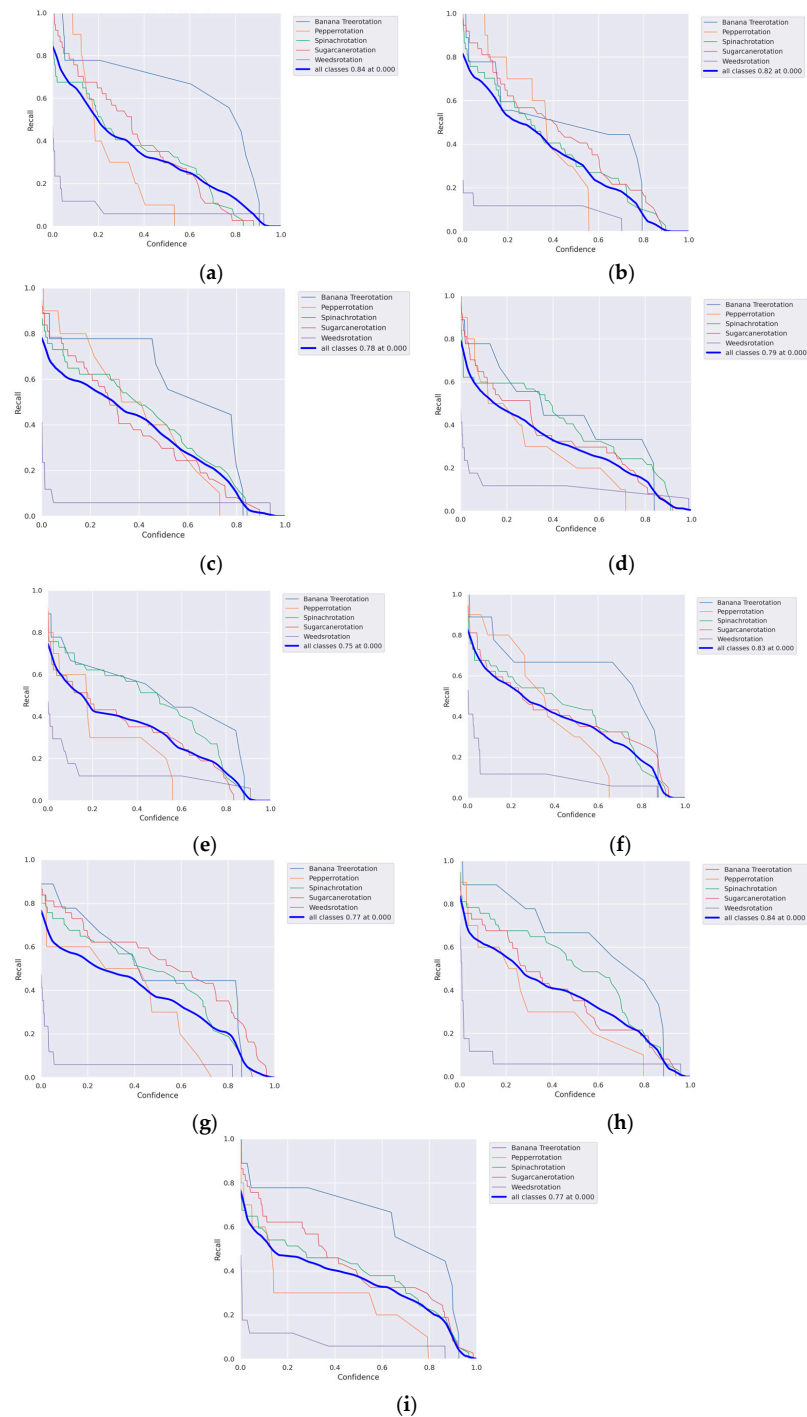

Figure 6. Recall confidence of batch size training and validation at (**a**) batch size 10, (**b**) batch size 20, (**c**) batch size 30, (**d**) batch size 40, (**e**) batch size 50, (**f**) batch size 60, (**g**) batch size 70, (**h**) batch size 80, and (**i**) batch size 90.

The confusion matrices obtained for all the batch sizes are presented in Figure 7a–i. In the confusion matrices, the diagonal elements, running from the top left to the bottom right represent the number of true positive (TP) predictions for each class. Figure 7a presents the confusion matrix for batch size 10. Specifically, it shows that 67% of items in the "banana" class, 30% in the "pepper" class, 46% in the "spinach" class, 59% in the "sugarcane" class, and 6% in the "weed" class were correctly classified.

Conversely, 33%, 70%, 54%, 41%, and 94% of objects belonging to the banana class, pepper class, spinach class, sugarcane class, and weed class, respectively, were classified as "unknown." These are instances where the classifier could not confidently assign these objects to any specific class.
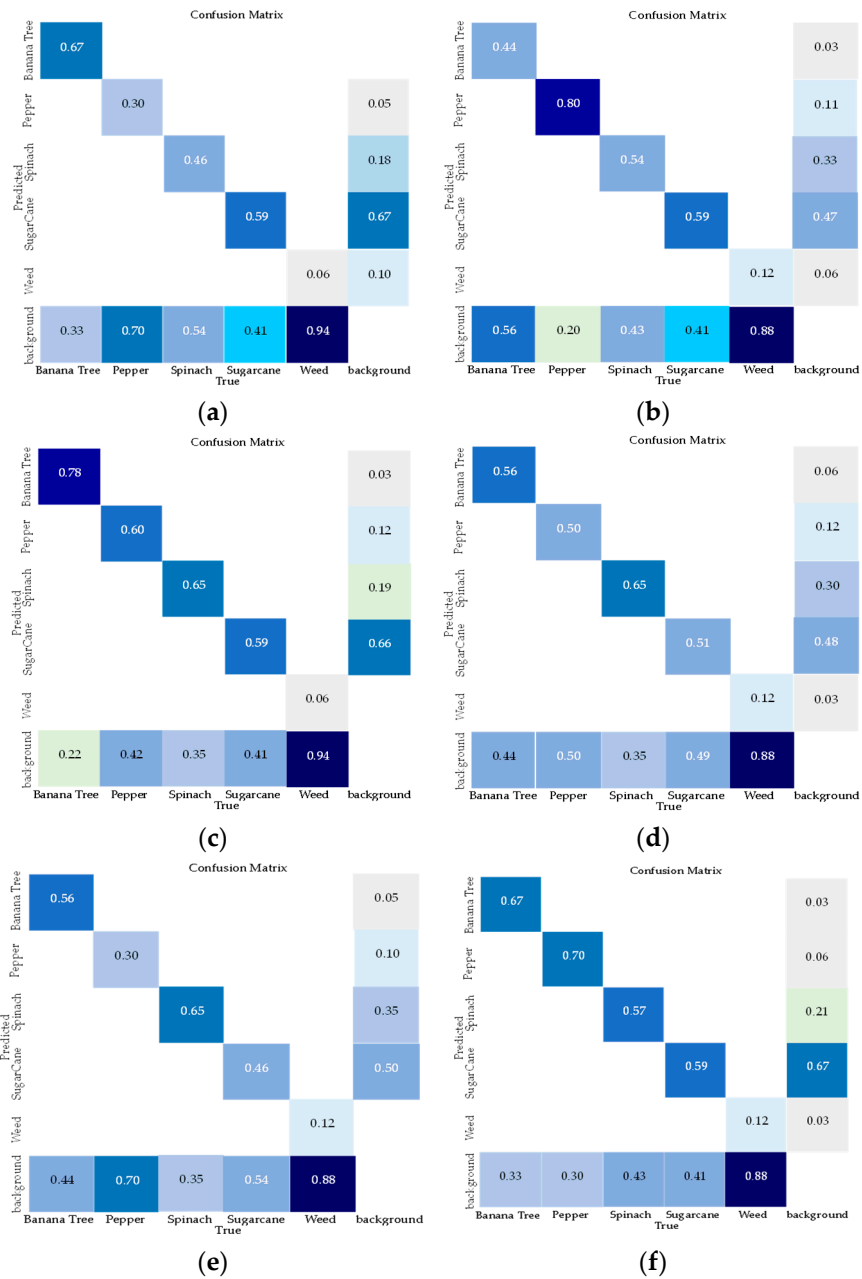

Figure 7. *Cont.*
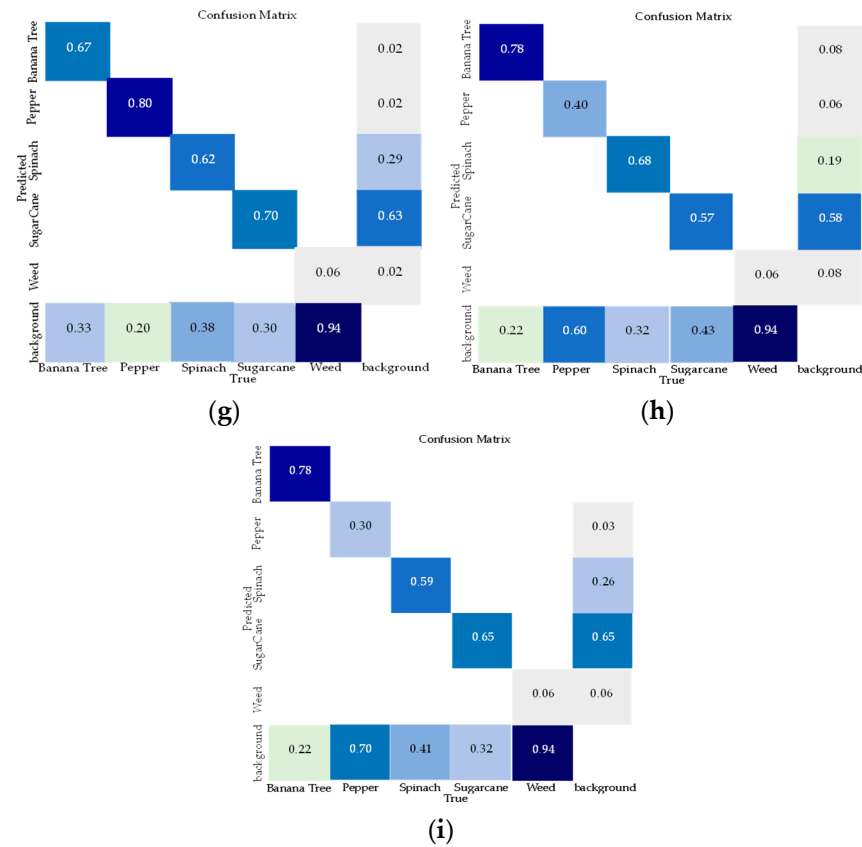
(g)

(h)

(i)

Figure 7. Confusion Matrix for (**a**) batch size 10, (**b**) batch size 20, (**c**) batch size 30, (**d**) batch size 40, (**e**) batch size 50, (**f**) batch size 60, (**g**) batch size 70, (**h**) batch size 80, and (**i**) batch size 90.

Tables 1–9 present the crop-specific performance of the model at different batch sizes. In these tables, the precision values span from 0 (indicating no precision) to 1 (perfect precision), while the recall values also range from 0 (no recall) to 1.0 (ideal recall). For batch size 10 (see Table 1), among the different classes 'banana' exhibited the highest precision at approximately 0.884, making it the most precise class. This was followed by 'spinach', 'pepper crops', 'sugarcane crops', and finally 'weeds' with approximately 0.278 precision, ranking as the least precise. For recall, the 'banana' class achieved the highest recall value at approximately 0.778, making it the best recognized class. In descending order, the classes 'sugarcane crops', 'spinach', 'pepper crops', and 'weeds' followed, with 'weeds' having the lowest recall value at approximately 0.118. This suggests that the classifier identified fewer positive samples of 'weeds' compared to 'spinach', 'bananas', 'pepper crops', and 'sugarcane'.

Table 1. Details of the precision and recall obtained at batch size 10.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.884	0.778	0.874	0.553
Pepper	52	10	0.536	0.578	0.577	0.226
Spinach	52	37	0.561	0.595	0.554	0.219
Sugarcane	52	37	0.29	0.676	0.431	0.15
Weeds	52	17	0.278	0.118	0.0962	0.0422

Table 2. Details of the precision and recall obtained at batch size 20.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.826	0.530	0.694	0.436
Pepper	52	10	0.612	0.634	0.733	0.275
Spinach	52	37	0.513	0.486	0.508	0.226
Sugarcane	52	37	0.511	0.568	0.509	0.202
Weeds	52	17	0.528	0.118	0.121	0.0703

Table 3. Details of the precision and recall obtained at batch size 30.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.882	0.778	0.833	0.571
Pepper	52	10	0.503	0.609	0.695	0.237
Spinach	52	37	0.669	0.622	0.617	0.246
Sugarcane	52	37	0.439	0.529	0.445	0.195
Weeds	52	17	0.485	0.0588	0.0923	0.0437

Table 4. Details of the precision and recall obtained at batch size 40.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.635	0.582	0.712	0.483
Pepper	52	10	0.423	0.442	0.441	0.135
Spinach	52	37	0.521	0.595	0.538	0.262
Sugarcane	52	37	0.304	0.514	0.368	0.152
Weeds	52	17	0.496	0.118	0.185	0.0807

Table 5. Details of the precision and recall obtained at batch size 50.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.857	0.665	0.777	0.540
Pepper	52	10	0.465	0.600	0.495	0.129
Spinach	52	37	0.519	0.703	0.660	0.268
Sugarcane	52	37	0.344	0.568	0.428	0.209
Weeds	52	17	0.292	0.176	0.201	0.0628

Table 6. Details of the precision and recall obtained at batch size 60.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.758	0.700	0.879	0.517
Pepper	52	10	0.796	0.800	0.762	0.190
Spinach	52	37	0.595	0.597	0.580	0.240
Sugarcane	52	37	0.306	0.568	0.500	0.264
Weeds	52	17	0.425	0.118	0.162	0.0532

Table 7. Details of the precision and recall obtained at batch size 70.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.870	0.745	0.818	0.524
Pepper	52	10	0.393	0.600	0.391	0.102
Spinach	52	37	0.555	0.649	0.593	0.234
Sugarcane	52	37	0.349	0.723	0.546	0.249
Weeds	52	17	0.172	0.0588	0.101	0.0509

Table 8. Details of the precision and recall obtained at batch size 80.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.717	0.847	0.878	0.561
Pepper	52	10	0.505	0.510	0.598	0.157
Spinach	52	37	0.543	0.676	0.648	0.25
Sugarcane	52	37	0.389	0.655	0.407	0.206
Weeds	52	17	0.211	0.0588	0.110	0.048

Table 9. Details of the precision and recall obtained at batch size 90.

Class	Images	Instances	Precision	Recall	mAP50	mAP50-95
Banana Tree	52	9	0.965	0.778	0.860	0.525
Pepper	52	10	0.397	0.300	0.423	0.117
Spinach	52	37	0.566	0.493	0.505	0.221
Sugarcane	52	37	0.497	0.614	0.475	0.196
Weeds	52	17	0.452	0.101	0.155	0.0601

The confusion matrix obtained at batch size 20 for the multi-class classification is illustrated in Figure 7b. Specifically, it shows that 44% of items in the "banana" class, 80% in the "pepper" class, 57% in the "spinach" class, 59% in the "sugarcane" class, and 6% in the "weed" class were correctly classified. Conversely, 56%, 20%, 43%, 41%, and 88% of objects belonging to the banana, pepper, spinach, sugarcane, and weed classes, respectively, were classified as "unknown."

As shown in Table 2, 'banana' also exhibited the highest precision at approximately 0.826, making it the most precise. This was followed by 'pepper crops', 'weeds', and 'spinach', in that order, with 'sugarcane' ranking as the least precisely detected with approximately 0.511 precision. Notably, the 'pepper' class achieved the highest recall, with an approximate value of 0.634. In descending order, 'sugarcane crops', banana', 'spinach', and finally 'weeds', with an approximate recall value of 0.118 followed, indicating that the classifier identified fewer positive instances of 'sugarcane', 'spinach', 'bananas', and 'pepper crops' and the least positive instances of 'weeds'.

The confusion matrix obtained at batch size 30 for the classification (see Figure 7c) shows that 78% of items in the "banana" class, 60% in the "pepper" class, 65% in the "spinach" class, 59% in the "sugarcane" class, and 6% in the "weed" class were correctly classified. Conversely, 22%, 40%, 35%, 41%, and 94% of objects belonging to the banana class, pepper class, spinach class, sugarcane class, and weed class, respectively, were classified as "unknown." These instances could not be confidently assigned to any specific class by the classifier.

Table 3 shows that 'banana' exhibited the highest precision at approximately 0.882, making it the most precise. This was followed by 'spinach', 'pepper', 'weeds', and finally 'sugarcane crops' with approximately 0.439 precision ranking as the least precise. In addition, the 'banana' class achieved the highest recall, with an approximate value of 0.778. In descending order, 'spinach', pepper', and 'sugarcane' followed, with 'weeds' returned the lowest recall with an approximate value of 0.0588, indicating that the classifier identified few or no positive instances.

Figure 7d presents the confusion matrix obtained at batch size 40 for the crop classification. Specifically, it shows that 56% of items in the "banana" class, 50% in the "pepper" class, 65% in the "spinach" class, 51% in the "sugarcane" class, and 12% in the "weed" class were correctly classified. Conversely, 44%, 50%, 35%, 49%, and 88% of objects belonging to the banana class, pepper class, spinach class, sugarcane class, and weed class, respectively, were classified as "unknown."

Table 4 also shows that 'banana' exhibited the highest precision at approximately 0.635, which was followed by 'spinach', 'weeds', 'pepper crops', and finally 'sugarcane crops' with approximately 0.304 precision ranking as the least precise. On the other hand, the 'spinach'

class achieved the highest recall, with an approximate value of 0.595. In descending order, 'banana', sugarcane crops', 'pepper crops', and 'weeds', with an approximate recall value of 0.118 followed, indicating that the classifier identified few positive instances of 'sugarcane', 'spinach', 'bananas', and 'pepper crops' and the least positive instances of 'weeds'.

The confusion matrix obtained at batch size 50 for the classification (see Figure 7e) shows that 56% of items in the "banana" class, 30% in the "pepper" class, 65% in the "spinach" class, 46% in the "sugarcane" class, and 12% in the "weeds" class were correctly classified. Conversely, 44%, 70%, 35%, 54%, and 88% of objects belonging to the banana class, pepper class, spinach class, sugarcane class, and weed class, respectively, were classified as "unknown."

Of all the classes presented in Table 5, 'banana' exhibited the highest precision at approximately 0.857, making it the most precise. This was followed by 'spinach', 'pepper crops', 'sugarcane', and finally 'weeds' with approximately 0.292 precision ranking as the least precise. The 'spinach' class also achieved the highest recall, with an approximate value of 0.703. In descending order, 'banana', pepper crops', 'sugarcane crops', and 'weeds' followed, the latter with an approximate recall value of 0.176, indicating that the classifier identified few or no positive instances.

Figure 7f presents the confusion matrix obtained at batch size 60 for the multi-class classification. It shows that 67% of items in the "banana" class, 70% in the "pepper" class, 57% in the "spinach" class, 59% in the "sugarcane" class, and 12% in the "weed" class were correctly classified. Conversely, 33%, 30%, 43%, 41%, and 88% of objects belonging to the banana class, pepper class, spinach class, sugarcane class, and weed class, respectively, were classified as "unknown".

Table 6 shows that 'pepper' exhibited the highest precision at approximately 0.796, making it the most precise. This was followed by 'banana', 'spinach', 'weed', and finally 'sugarcane crops' with approximately 0.306 precision ranking as the least precise. The 'pepper crops' class also achieved the highest recall, with an approximate value of 0.800, making it the class with the best recall. In descending order, 'banana', spinach', 'sugarcane crop', followed, while 'weeds', with an approximate recall value of 0.118, yielded the lowest recall.

The confusion matrix obtained at Batch size 70 for the crop classification is illustrated in Figure 7g, showing that 67% of items in the "banana" class, 80% in the "pepper" class, 62% in the "spinach" class, 70% in the "sugarcane" class, and 6% in the "weed" class were correctly classified. Conversely, 33%, 20%, 38%, 30%, and 94% of objects belonging to the banana class, pepper class, spinach class, sugarcane class, and weed class, respectively, were classified as "unknown".

As shown in Table 7, The 'banana' class exhibited the highest precision at approximately 0.870, making it the most precise. This was followed by 'spinach', 'pepper', 'sugarcane crops', and finally 'weeds' with approximately 0.172 precision ranking as the least precise. Notably, the 'banana' class also achieved the highest recall, with an approximate value of 0.745, making it the class with the best recall. In descending order, 'sugarcane crop', spinach', 'pepper crop', and 'weeds' followed, the latter with an approximate recall value of 0.0588, indicating that the classifier identified fewer positive instances of 'sugarcane', 'spinach', 'bananas', and 'pepper crops' and the least positive instances of 'weeds'.

The confusion matrix obtained at batch size 80 for the multi-crop classification (see Figure 7h) shows that 78% of items in the "banana" class, 40% in the "pepper" class, 68% in the "spinach" class, 57% in the "sugarcane" class, and 6% in the "weeds" class were correctly classified. Conversely, 22%, 60%, 32%, 43%, and 94% of objects belonging to the banana class, pepper class, spinach class, sugarcane class, and weeds class, respectively, were classified as "unknown." In Table 8, it can be observed that 'banana' exhibits the highest precision at approximately 0.717, making it the most precise. This is followed by 'spinach', 'pepper', 'sugarcane crops', and finally 'weeds' with approximately 0.211 precision ranking as the least precise. The 'banana' class also achieves the highest recall, with an approximate value of 0.847, followed in descending order by 'spinach', 'sugarcane

crops', 'pepper crops', and finally 'weeds' with an approximate recall value of 0.0588, indicating that the classifier identified fewer positive instances of 'sugarcane', 'spinach', 'bananas', and 'pepper crops' and the least positive instances of 'weeds' at this batch size.

Figure 7i presents the confusion matrix obtained at batch size 90 for the multi-class classification. It shows that 78% of items in the "banana" class, 30% in the "pepper" class, 59% in the "spinach" class, 65% in the "sugarcane" class, and 6% in the "weed" class were correctly classified. Conversely, 22%, 70%, 41%, 35%, and 94% of objects belonging to the banana class, pepper class, spinach class, sugarcane class, and weed class, respectively, were classified as "unknown." As shown in Table 9, 'banana' exhibited the highest precision at approximately 0.965, followed by 'spinach', 'sugarcane', 'weeds', and then 'pepper' with approximately 0.397 precision. Likewise, the 'banana' class achieved the highest recall, with an approximate value of 0.778. In descending order, 'sugarcane crops', 'spinach', 'pepper crops', and 'weeds' followed, indicating that the classifier identified fewer positive instances of 'sugarcane', 'spinach', 'bananas', and 'pepper crops' and the least positive instances of 'weeds'.

3.2. Overall Model Performance

The overall accuracy, precision, recall, and F1 score results recorded at each batch size, depicting the overall performance of the model in automatic crop classification, are presented in Table 10.

Table 10. Overall performance of the model (non-crop-specific).

Batch Size	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
10	50.60	95	84	18
20	51.30	84	82	34
30	53.60	96	78	29
40	44.90	87	79	20
50	51.20	82	75	8
60	57.70	89	83	18
70	49.00	93	77	17
80	52.80	98	84	23
90	48.40	95	82	25

4. Discussion

Throughout the tested batch sizes, 'banana' exhibited superior performance according to most of the assessment metrics. In contrast, 'sugarcane' and 'weeds' showed relatively low precision rates, while 'spinach' and 'pepper' yielded average precision. Notably, at batch size 20 all crops returned above average precision. Additionally, the best precision for 'pepper' was recorded at batch size 60. The impressive performance of the model in automatic classification of the banana class, despite using the same quantity and quality of training data as other classes, can be attributed to several factors. First, the distinct visual features of banana leaves, such as their shape, size, and texture, set them apart from other plants on the farm. Although the colors of the other plants, including weeds, are similar, these unique visual characteristics make it easier for the model to distinguish bananas from other classes. Additionally, the banana class exhibited less variability within itself, meaning that the appearances of bananas in the images were more consistent. This consistency allowed the model to learn and recognize them more accurately. In addition, the banana plants were likely less obstructed by other objects or plants, making their features more visible. This clear visibility aids in more accurate classification by the model. Furthermore, bananas tend to have a higher contrast with their background and neighboring crops, which makes their features stand out more prominently in the images. This high contrast enhances the model's ability to detect and classify them accurately.

The poor performance of the model in terms of precision when classifying certain classes, such as 'weeds' and 'sugarcane', can be attributed to several factors. One significant

issue is the lack of distinct visual features in these classes compared to others such as bananas. In particular, weeds tend to have a wide variety of appearances, making it difficult for the model to learn a consistent set of characteristics for accurate identification. This high intra-class variability means that weeds can look very different from one another, leading to more classification errors. Another factor is the similarity in color and texture between weeds and the other crops in the dataset. This similarity can confuse the model, as it struggles to differentiate between weeds and certain crops, especially in mixed-crop environments. These overlapping visual features lead to lower precision, as the model incorrectly identifies non-weed objects as weeds and vice versa.

For sugarcane, the performance issues could stem from the physical characteristics and growing patterns of the plant. Sugarcane plants are often tall and densely packed, which can result in significant occlusion. This occlusion means that parts of the sugarcane plants are blocked from view, preventing the model from seeing the complete structure and reducing the accuracy of classification. Additionally, the repetitive and similar appearance of sugarcane stalks can make it challenging for the model to identify unique features that distinguish sugarcane from other classes.

Furthermore, spinach leaves are typically smaller and can be more variable in shape and size, which can make it difficult for the model to learn a consistent set of features by which to identify them accurately. Spinach often grows close to the ground and can be covered by other vegetation, resulting in occlusion that makes it challenging for the model to acquire a clear view of the entire plant. Similarly, pepper plants often have leaves and fruits that blend in with the surrounding foliage, making it harder for the model to distinguish them from other plants or background elements.

The overall accuracy obtained for batch sizes 10, 20, 30, 40, 50, 60, 70, 80, and 90 was 50%, 51%, 54%, 44%, 51%, 58%, 49%, 53%, and 48%, respectively, as shown in Table 10. This trend indicates a steady improvement in accuracy from batch sizes 10 to 30, a dip at batch size 40, and a peak at batch size 60, followed by a fluctuating pattern. Batch size 10 achieved average precision, recall, and F1 scores of 95%, 84%, and 18%, respectively, while batch size 20 scored 84%, 82%, and 34%. At batch size 30, results of 96%, 78%, and 29% were obtained, whereas batch size 40 scored 87%, 79%, and 20%. For batch size 50, the metrics were 82%, 75%, and 8%, while batch size 60 recorded results of 89%, 83%, and 18%. Batch size 70 yielded results of 93%, 77%, and 17%, while batch 80 achieved scores of 98%, 84%, and 23%. Finally, batch size 90 scored 95%, 82%, and 25%. The accuracy of the predictions fluctuated up to batch size 40, then improved consistently from batch sizes 50 to 80, supporting the findings of [26,27], which affirmed that increased batch size enhances model accuracy. This is also corroborated by [30,31], which used YOLOv8 for vegetable disease and wheat seed detection, respectively.

The marginal increase in accuracy from batch size 10 (50.6%) to batch size 20 (51.3%) suggests a positive trend. However, the average precision drops significantly from 95% to 84%, implying a higher likelihood of false positives at batch size 20 despite only a modest decrease in recall from 84% to 82%. The F1 score improved notably from 18% to 34%, indicating better balance in minimizing false positives and negatives. Batch size 30 showed a slight improvement in accuracy (53.6%) over batch size 20 (51.3%). The precision increased significantly from 84% to 96%, while the recall decreased from 82% to 78%, suggesting more false negatives. The F1 score decreased from 34% to 29%, reflecting this trade-off.

At batch size 40, the accuracy dropped to 44.9%, with precision at 87% and recall at 79%. The F1 score further decreased to 20%, indicating a less optimal balance. Batch size 50 saw the accuracy rise to 51.2%, although precision and recall dropped to 82% and 75%, respectively, with the F1 score plummeting to 8%. Batch size 60 marked an accuracy peak at 57.7%, with precision and recall at 89% and 83%, respectively. The F1 score improved to 18%, indicating better balance. However, the accuracy at batch size 70 fell to 49% despite high precision (93%) and lower recall (77%), resulting in an F1 score of 17%.

Batch size 80 saw improved accuracy of 52.8%, with exceptional precision (98%) and higher recall (84%), yielding an F1 score of 23%. This batch size achieved a commendable balance, which is crucial for minimizing false positives and negatives. Finally, the accuracy at batch size 90 decreased to 48.4%, with precision, recall, and F1 scores of 95%, 82%, and 25%, respectively, indicating a slight overall improvement in balance despite the lower accuracy.

The initial transition from batch sizes 10 to 60 suggests a consistent trend of increasing accuracy, which indicates that the model benefits from larger batches during training up to a certain point. This progression implies that larger batch sizes contribute positively to the model's ability to identify and classify crop types accurately. However, at batch size 70 a sudden dip in accuracy disrupts the upward trend, prompting a re-evaluation of the relationship between batch size and accuracy. This finding challenges the conventional expectation that increasing batch sizes consistently lead to improved model performance [21,22,24]. The subsequent batch sizes of 80 and 90 showed a resurgence in accuracy, introducing a sinusoidal-like oscillation to the results.

In general, this study demonstrates that increasing batch size does not always lead to better performance. Batch size 30 showed notable improvements in certain metrics, but optimal performance was observed at batch sizes 60 and 80 with a more balanced high precision and recall, which aligns with the findings of [32]. The pattern of the results from our model suggests that factors beyond batch size alone influence the YOLOv8 model's performance. The characteristics of the dataset, neural network architecture, and specifics of the training process all contribute to this multifaceted dynamic. These variations highlight the model's sensitivity to hyperparameter tuning and the importance of carefully tuning batch sizes based on the specific objectives and requirements of each project. Additionally, trade-offs between evaluation metrics (accuracy, precision, recall, and F1 score) should be considered when selecting the most suitable batch size for crop classification tasks.

5. Conclusions

This study evaluated the performance of YOLOv8, a Convolutional Neural Network (CNN) model, for automatic crop classification in a mixed crop farmland using drone-acquired images. The model's performance was systematically assessed across various training batch sizes using diverse metrics, including loss function graphs, precision and recall graphs, confusion matrices, and validation metrics such as F1 score, accuracy, recall, and precision.

Based on our analysis, batch size 60 stands out with the highest accuracy, indicating that this batch size provides the best overall performance for automatic crop detection and classification. Although the F1 score is moderate, the combination of high accuracy, precision, and recall makes it the most balanced option. However, batch size 80 also shows very high precision (98%) and balanced recall (84%), which might be suitable if the primary focus is on achieving high precision. This implies that for optimal results in terms of balanced performance across all metrics, batch size 60 is recommended, while if precision is the primary concern, then batch size 80 would be a suitable alternative.

In summary, YOLOv8 maintains comparable detection accuracy in identifying and classifying crops within a mixed crop farmland. Given the limited dataset of approximately 393 image pairs used in this study, future research endeavors should explore the model's performance with a more extensive collection of crop images to better evaluate the model's capabilities and limitations in diverse agricultural scenarios.

Author Contributions: Conceptualization; O.G.A.; methodology, O.G.A. and O.S.A.; software and visualization, O.S.A.; formal analysis, O.S.A. and P.O.I.; resources, O.G.A.; writing—original draft preparation, O.S.A.; writing—review & editing, O.G.A. and P.O.I.; supervision, O.G.A. and P.O.I.; project administration, O.G.A. and P.O.I. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data and any other data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gray, H.; Nuri, K.R. Differing visions of agriculture: Industrial-chemical vs. small farm and urban organic production. *Am. J. Econ. Sociol.* **2020**, *79*, 813–832. [CrossRef]
2. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A.M. Deep learning techniques to classify agricultural crops through UAV imagery: A review. *Neural Comput. Applic* **2022**, *34*, 9511–9536. [CrossRef] [PubMed]
3. Hall, O.; Dahlin, S.; Marstorp, H.; Archila Bustos, M.F.; Öborn, I.; Jirström, M. Classification of Maize in Complex Smallholder Farming Systems Using UAV Imagery. *Drones* **2018**, *2*, 22. [CrossRef]
4. Dhanaraju, M.; Chenniappan, P.; Ramalingam, K.; Pazhanivelan, S.; Kaliaperumal, R. Smart Farming: Internet of Things (IoT)-Based Sustainable Agriculture. *Agriculture* **2022**, *12*, 1745. [CrossRef]
5. Sorenson, C.J. Crop Classification in Mixed Farming Systems. *J. Agric. Sci.* **2007**, *145*, 469–480.
6. He, S.; Peng, P.; Chen, Y.; Wang, X. Multi-Crop Classification Using Feature Selection-Coupled Machine Learning Classifiers Based on Spectral, Textural and Environmental Features. *Remote Sens.* **2022**, *14*, 3153. [CrossRef]
7. Bhosle, K.; Musande, V. Evaluation of CNN model by comparing with convolutional autoencoder and deep neural network for crop classification on hyperspectral imagery. *Geocarto Int.* **2020**, *37*, 813–827. [CrossRef]
8. Rodriguez, M. Precision Agriculture: A New Era in Farming. *J. Agric. Sci.* **2015**, *153*, 171–182.
9. Feng, Q.; Yang, J.; Liu, Y.; Ou, C.; Zhu, D.; Niu, B.; Liu, J.; Li, B. Multi-temporal unmanned aerial vehicle remote sensing for vegetable mapping using an attention-based recurrent convolutional neural network. *Remote Sens.* **2020**, *12*, 1668. [CrossRef]
10. Monteiro, A.; Santos, S.; Gonçalves, P. Precision Agriculture for Crop and Livestock Farming—Brief Review. *Animals* **2021**, *11*, 2345. [CrossRef]
11. Csillik, O.; Cherbini, J.; Johnson, R.; Lyons, A.; Kelly, M. Identification of citrus trees from unmanned aerial vehicle imagery using convolutional neural networks. *Drones* **2018**, *2*, 39. [CrossRef]
12. Siesto, G.; Fernández-Sellers, M.; Lozano-Tello, A. Crop Classification of Satellite Imagery Using Synthetic Multitemporal and Multispectral Images in Convolutional Neural Networks. *Remote Sens.* **2021**, *13*, 3378. [CrossRef]
13. Somching, N.; Wongsai, S.; Wongsai, N.; Koedsin, W. Using Machine Learning Algorithm and Landsat Time Series to Identify Establishment Year of Para Rubber Plantations: A Case Study in Thalang District, Phuket Island, Thailand. *Int. J. Remote Sens.* **2020**, *41*, 9075–9100. [CrossRef]
14. Patel, R. Crop Classification Using Deep Learning. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 1–6.
15. Ajayi, O.G.; Ashi, J.; Guda, B. Performance evaluation of YOLO v5 model for automatic crop and weed classification on UAV images. *Smart Agric. Technol.* **2023**, *5*, 100231. [CrossRef]
16. Zhang, J.; Wang, T.; Wang, B.; Chen, C.; Wang, G. Hyperparameter optimization method based on dynamic Bayesian with sliding balance mechanism in neural network for cloud computing. *J. Cloud Comp.* **2023**, *12*, 109. [CrossRef]
17. Radiuk, P.M. Impact of training set batch size on the performance of convolutional neural networks for diverse datasets. *Inf. Technol. Manag. Sci.* **2017**, *20*, 20–24. [CrossRef]
18. Keskar, N.S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; Tang, P.T.P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In Proceedings of the 2016 International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
19. Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; He, K. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv* **2017**, arXiv:1706.02677.
20. Smith, S.L.; Kindermans, P.J.; Ying, C.; Le, Q.V. Don't Decay the Learning Rate, Increase the Batch Size. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018; Available online: https://openreview.net/pdf?id=B1Yy1BxCZ (accessed on 15 February 2024).
21. You, Y.; Gitman, I.; Ginsburg, B. Large Batch Training of Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5477–5486.
22. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
23. Ajayi, O.G.; Opaluwa, Y.D.; Ashi, J.; Zikirullahi, W.M. Applicability of artificial neural network for automatic crop type classification on UAV-based images. *Environ. Technol. Sci. J.* **2022**, *13*, 57–72. [CrossRef]
24. Ajayi, O.G.; Ashi, J. Effects of Varying Training Epochs of a Faster Region-Based Convolutional Neural Network on the Accuracy of an Automatic Weed Classification Scheme. *Smart Agric. Technol.* **2023**, *3*, 100128. [CrossRef]
25. Kamilaris, A.; Prenafeta-Boldú, F.X. A review of the use of convolutional neural networks in agriculture. *J. Agric. Sci.* **2018**, *156*, 312–322. [CrossRef]

26. Corceiro, A.; Alibabaei, K.; Assunção, E.; Gaspar, P.D.; Pereira, N. Methods for Detecting and Classifying Weeds, Diseases and Fruits Using AI to Improve the Sustainability of Agricultural Crops: A Review. *Processes* **2023**, *11*, 1263. [CrossRef]
27. Qu, H.-R.; Su, W.-H. Deep Learning-Based Weed–Crop Recognition for Smart Agricultural Equipment: A Review. *Agronomy* **2024**, *14*, 363. [CrossRef]
28. Shen, L.; Lang, A.; Songl, Z. Infrared Object Detection Method Based on DBD-YOLOv8. *IEEE Access* **2023**, *11*, 145853–145868. [CrossRef]
29. Ajayi, O.G.; Olufade, O.O. Drone-based crop type identification with convolutional neural networks: An evaluation of the performance of RESNET architectures. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *X-1/W1-2023*, 991–998. [CrossRef]
30. Wang, X.; Liu, J. Vegetable disease detection using an improved YOLOv8 algorithm in the greenhouse plant environment. *Sci. Rep.* **2024**, *14*, 4261. [CrossRef]
31. Ma, N.; Su, Y.; Yang, L.; Li, Z.; Yan, H. Wheat Seed Detection and Counting Method Based on Improved YOLOv8 Model. *Sensors* **2024**, *24*, 1654. [CrossRef]
32. Ajayi, O.G.; Iwendi, E.; Adetunji, O.O. Optimizing crop classification in precision agriculture using AlexNet and high resolution UAV imagery. *Technol. Agron.* **2024**, *4*, e011. [CrossRef]