TETHYS ENVIRONMENTAL SCIENCE
an International Journal

Comparative Evaluation of YOLO-Based Pretrained Models for Multi-Class Vegetable Detection

Sakibul Hasan Chowdhury[1], Yakup Kutlu[2]*, Nalanda Dewan Goongoon[3]

[1]Department of Software Engineering, Daffodil International University, Data Science Laboratory, Daffodil Smart City, 1216 Dhaka, Bangladesh.
[2]Department of Computer Engineering, Iskenderun Technical University, 31220 Iskenderun, Hatay, Türkiye.
[3]Department of Information Technology and Management, Daffodil International University, Daffodil Smart City, 1216 Dhaka, Bangladesh.

Research Article

Citation: Chowdhury, S. H., Kutlu, Y., Goongoon, N. D. (2026). Comparative Evaluation of YOLO-Based Pretrained Models for Multi-Class Vegetable Detection. *Tethys Env. Sci.* 3(1): 16-30.

DOI: 10.5281/zenodo.19250039

Received: 20 January 2026

Accepted: 11 March 2026

Available Online: 27 March 2026

Publication Date: 29 June 2026

Abstract

Accurate detection of vegetables from images is an essential task in precision agriculture, automated food supply chains and smart retail systems. Even with a little labeled data, recent developments in deep learning, especially pretrained object detection models, have greatly enhanced performance on visual recognition tasks.. However, the relative effectiveness of modern pretrained object detection architectures on agricultural datasets remains underexplored. In this study, we present a comprehensive comparative evaluation of multiple state-of-the-art pretrained object detection models on a publicly available vegetable object detection dataset from Bangladesh. The evaluated models include RF-DETR (Large), YOLOv11 variants, Roboflow 3.0 Object Detection models and YOLOv12 variants with different capacity configurations (Fast, Accurate and Extra Large). Standard metrics like recall, precision and mAP@50 on a predetermined test set are used to evaluate performance. Experimental results demonstrate that YOLOv12 (Extra Large) significantly outperforms other models, achieving an mAP@50 of 79.0%, precision of 81.8% and recall of 80.7%. When implementing pretrained detectors in agricultural computer vision applications, the results emphasize the significance of model architecture and scale.

Keywords: *Vegetable detection, deep learning, object detection, YOLO, precision agriculture, computer vision.*

Introduction

Agriculture plays a vital role in the global economy, particularly in developing countries such as Bangladesh, where vegetable production contributes significantly to food security and rural

Correspondence: Yakup Kutlu
yakup.kutlu@iste.edu.tr

livelihoods. Automated visual recognition systems can support a wide range of agricultural applications, including yield estimation, automated harvesting, quality inspection and intelligent market analytics. However, traditional computer vision approaches relied heavily on handcrafted features, which often failed to generalize under challenging real-world conditions such as variable illumination, occlusion and complex backgrounds commonly encountered in agricultural environments.

Recent advances in deep learning have substantially improved visual recognition and object detection performance. Region-based detectors such as Faster R-CNN (Ren et al., 2015) and single-stage detectors such as SSD (Liu et al., 2016) and the YOLO (You Only Look Once) family (Redmon et al., 2016; Jocher et al., 2023) have enabled accurate and efficient object detection in complex scenes. Furthermore, transfer learning using models pretrained on large-scale datasets such as COCO has significantly improved detection performance even when only limited domain-specific datasets are available (Lin et al., 2014).

Deep learning–based detection approaches have been widely applied in agricultural vision tasks including fruit detection, crop disease identification and weed recognition. In particular, YOLO-based architectures have gained popularity due to their balance between detection accuracy and real-time inference capability. Several studies have demonstrated the effectiveness of YOLO models for fruit and vegetable detection under natural field conditions (Li et al., 2022; Khanna et al., 2024; Wang and Liu, 2024; Zhu et al., 2026). In addition to convolutional neural network (CNN)–based detectors, transformer-based architectures such as DETR have recently attracted attention due to their ability to capture global contextual relationships within images through attention mechanisms (Sakai et al., 2016; Zeng, 2017; Carion et al., 2020). Hybrid approaches combining convolutional backbones with transformer-based decoders, such as RF-DETR, have further improved localization accuracy and robustness in complex scenes containing overlapping objects.

Recent research has also explored advanced deep learning models and datasets for agricultural object detection tasks. For example, Tapia-Mendez et al. (2025) evaluated several state-of-the-art object detectors on a fruit and vegetable dataset annotated with quality categories such as unripe, ripe and overripe, reporting strong performance for transformer-enhanced models such as DINO. In another study, Khanna et al. (2024) introduced the FRUVEG67 dataset and proposed the FVDNet architecture for fruit and vegetable detection in unconstrained environments, achieving competitive performance compared to recent YOLO variants. Similarly, Li et al. (2022) proposed an improved YOLOv5-based approach for vegetable disease detection, achieving high detection accuracy on a multi-scene dataset, while Wang and Liu (2024) developed an enhanced YOLOv8-based model incorporating attention mechanisms to improve detection performance in greenhouse environments.

Despite these advances, many existing studies focus on evaluating a single model architecture or dataset, making direct comparison between detection frameworks difficult. Systematic benchmarking studies that evaluate multiple modern pretrained detectors under consistent experimental settings remain relatively limited, particularly for vegetable detection datasets collected under real agricultural conditions (Camgözlü and Kutlu, 2020; Atasoy and Kutlu, 2022; Alamsyah et al., 2023; Camgözlü and Kutlu, 2023). Such comparative analyses are essential for understanding how different architectural designs perform in practical agricultural environments.

Therefore, this study presents a comprehensive evaluation of several contemporary pretrained object detection models using a Bangladeshi vegetable object detection dataset. The models are compared under a unified evaluation protocol using consistent performance metrics. The contributions of this study include a systematic comparison of modern pretrained object detectors on a vegetable detection dataset, quantitative evaluation using standardized metrics such as mAP@50, precision and recall, analysis of the influence of model scale and architecture on detection performance, and practical insights into the selection of object detection models for agricultural vision applications.

Materials and Methods

Dataset description

This study used the Vegetable Object Detection Dataset from Bangladesh (Jahan et al., 2025), publicly available on the Mendeley Data repository. The dataset contains annotated images of eight vegetable classes (beetroot, bitter gourd, bottle gourd, cabbage, capsicum, carrot, cauliflower and corn) captured under natural conditions. Images vary in illumination, scale, viewpoint and background complexity, making them suitable for real-world detection evaluation. Following the original distribution, the dataset was split into training, validation and test subsets, and all reported results were obtained from the held-out test set.

Preprocessing and annotation format

Prior to model training, all images were resized and normalized according to the default preprocessing requirements of each pretrained model architecture. Bounding box annotations were converted into the appropriate format required by the training and evaluation pipelines. No class rebalancing or oversampling procedures were applied in order to preserve the natural distribution of object classes within the dataset.

Pretrained object detection models

Several modern pretrained object detection models were evaluated, including RF-DETR (Large), YOLOv11 Extra Large, Roboflow 3.0 models (Accurate, Large, Extra Large) and YOLOv12 variants (Fast, Accurate, Extra Large). RF-DETR represents a transformer-based detector, while YOLOv11 and YOLOv12 are next-generation YOLO architectures designed for improved detection accuracy and efficiency. Roboflow models were included as baseline detectors with different capacity levels. All models were initialized with pretrained weights and fine-tuned on the vegetable dataset under identical experimental conditions.

Evaluation metrics

Model performance was evaluated using mean Average Precision at an Intersection over Union threshold of 0.5 (mAP@50), precision and recall. The mAP@50 metric measures detection accuracy across classes based on bounding box overlap, while precision and recall indicate the proportions of correct predictions relative to predicted and ground-truth objects. Together, these metrics provide a comprehensive assessment of detection accuracy, including localization performance, false positives and missed detections.

Experimental setup

During training, all images were resized to 512 × 512 pixels. Models were trained for 150 epochs with a batch size of 16 using the Adam optimizer with an initial learning rate of 0.001, and early stopping based on validation performance was applied to prevent overfitting. Transfer learning was used by initializing all models with pretrained weights from the COCO dataset. Data augmentation included horizontal flipping and saturation adjustment (−50% to +50%). All experiments were performed on the Roboflow 3.0 training platform using GPU-based cloud infrastructure to ensure consistent computational conditions across models.

Results and Discussion

RF-DETR (Large)

RF-DETR (Large) is a transformer-based object detection model that integrates convolutional feature extraction with global attention mechanisms, enabling robust object localization under complex backgrounds. Figure 1 shows that the training curves demonstrate a stable convergence with Box Location Loss, Class Loss, and Box Overlap Loss.

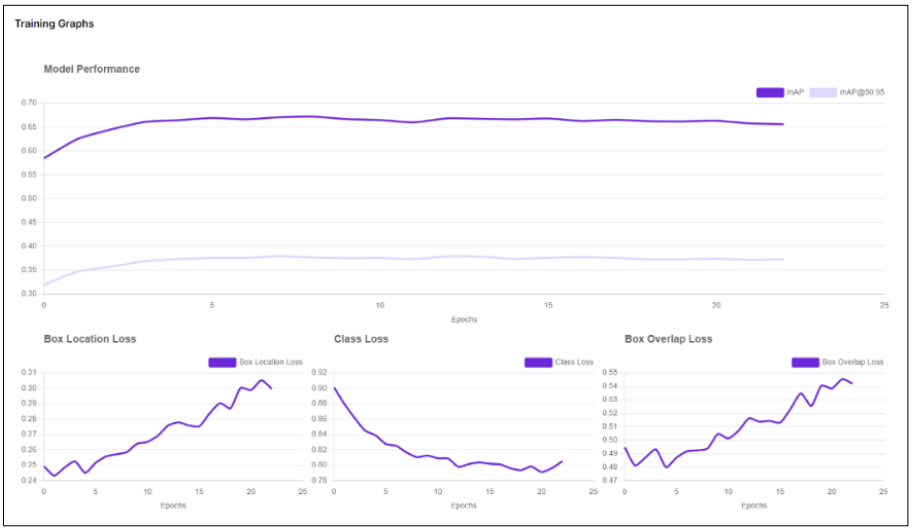

Figure 1. Training Graphs of RF-DETR (Large).

The model achieved an mAP@50 of 67.6%, with balanced precision (67.8%) and recall (67.3%). The Average Precision by class (Figure 2) shows strong detection performance for larger vegetables such as Pumpkin and bottle gourd, while comparatively lower AP scores were observed for smaller or visually similar classes such as capsicum and carrot. Overall, RF-DETR provides a strong baseline but is outperformed by newer YOLO architectures.

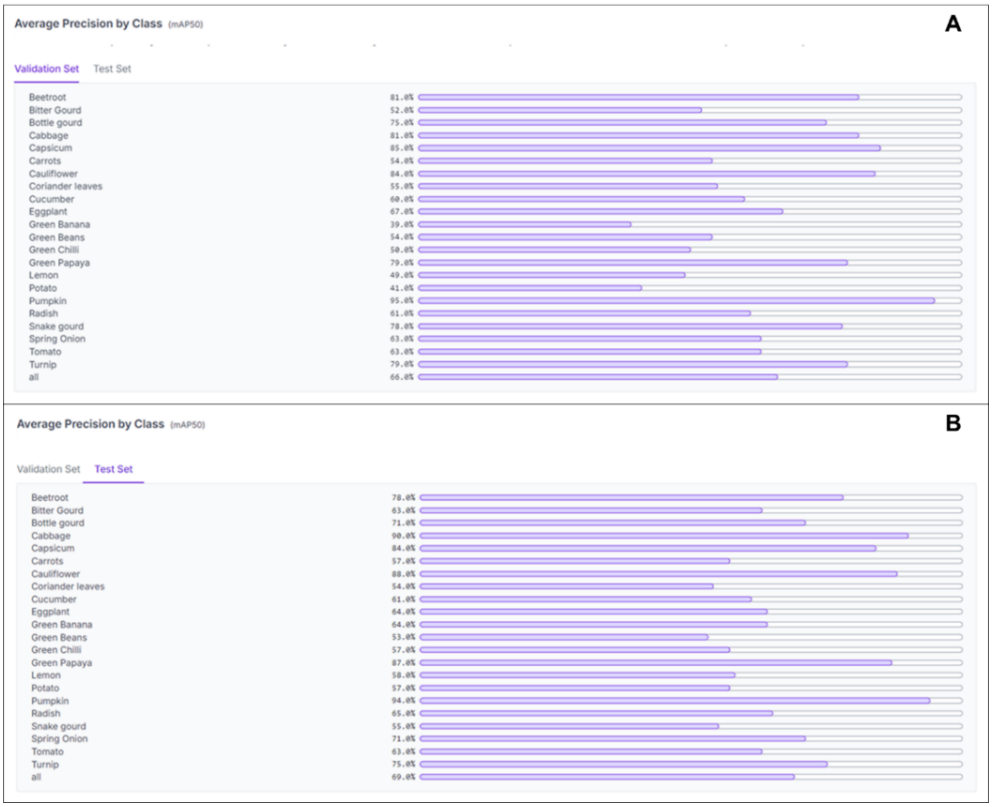

Figure 2. Class-wise average precision of the RF-DETR (Large) model on the (A) validation set and (B) test set.

YOLOv11 (Extra Large)

A second YOLOv11 (Extra Large) configuration was trained under identical conditions to evaluate training stability. As shown in Figure 3, training curves exhibit minor fluctuations in validation mAP, indicating sensitivity to initialization and optimization dynamics.

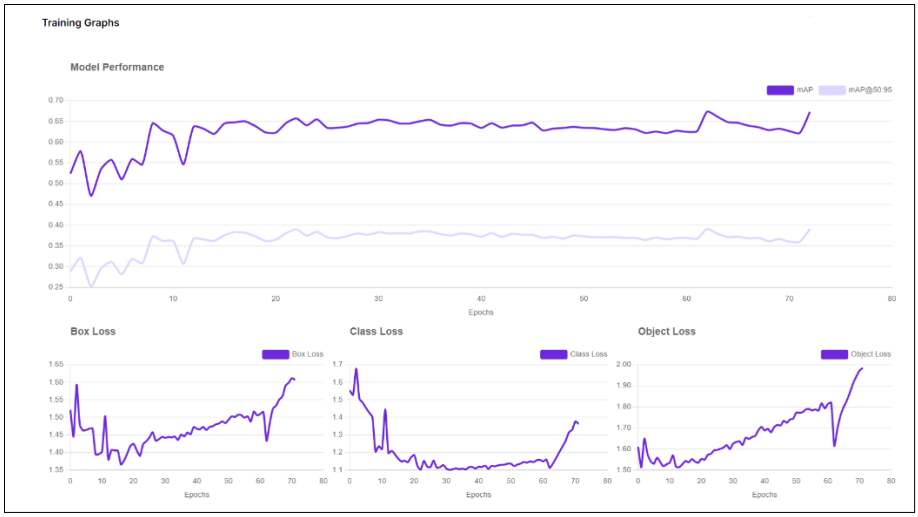

Figure 3. Training Graphs of YOLOv11 (Extra Large).

This variant achieved an mAP@50 of 59.8%, with a higher precision of 69.3% but slightly reduced recall (65.7%). The Average Precision by class (Figure 4) demonstrates marginal

improvement for certain classes, yet overall performance remains below transformer-based and YOLOv12 models.

Figure 4. Class-wise average precision of the YOLOv11 (Extra Large) model on the (A) validation set and (B) test set.

Roboflow 3.0 Object Detection (Accurate)

The Roboflow 3.0 Accurate model is designed to prioritize detection reliability over speed. As depicted in Figure 5, the training curves show steady learning behavior with minimal overfitting.

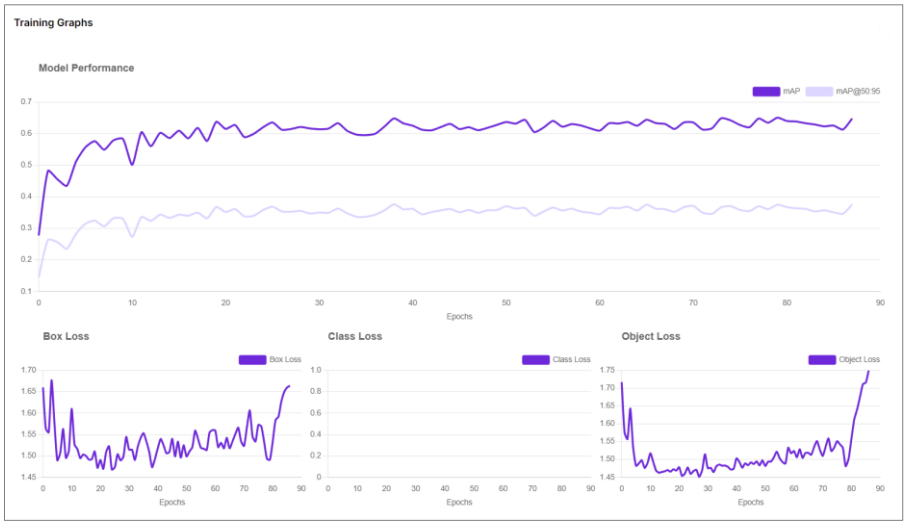

Figure 5. Training Graphs of Roboflow 3.0 Object Detection (Accurate).

The model obtained an mAP@50 of 58.8%, with both precision and recall at 65.9%. The class-wise AP plot (Figure 6) reveals uniform but moderate performance across most vegetable categories, indicating robustness but limited discriminative capacity for fine-grained class separation.

Figure 6. Class-wise average precision of the Roboflow 3.0 Object Detection (Accurate) model on the (A) validation set and (B) test set.

Roboflow 3.0 Object Detection (Extra Large)

The Extra Large Roboflow model further scales network depth and width. Figure 7, shows stable convergence with marginally improved mAP trends.

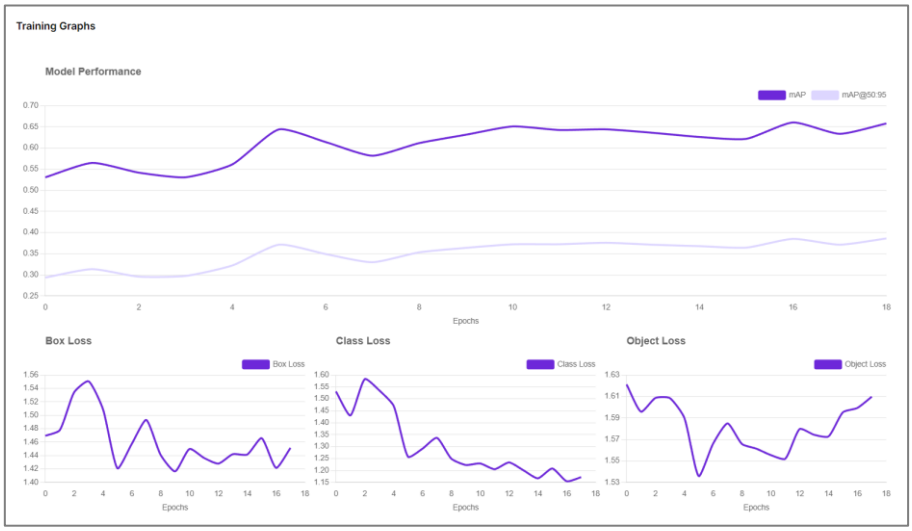

Figure 7. Training Graphs of Roboflow 3.0 Object Detection (Extra Large).

The model reached an mAP@50 of 59.4%, with precision of 67.5% and recall of 67.3%. The class-wise AP distribution (Figure 8) suggests diminishing returns from scaling within the Roboflow architecture compared to YOLOv12.

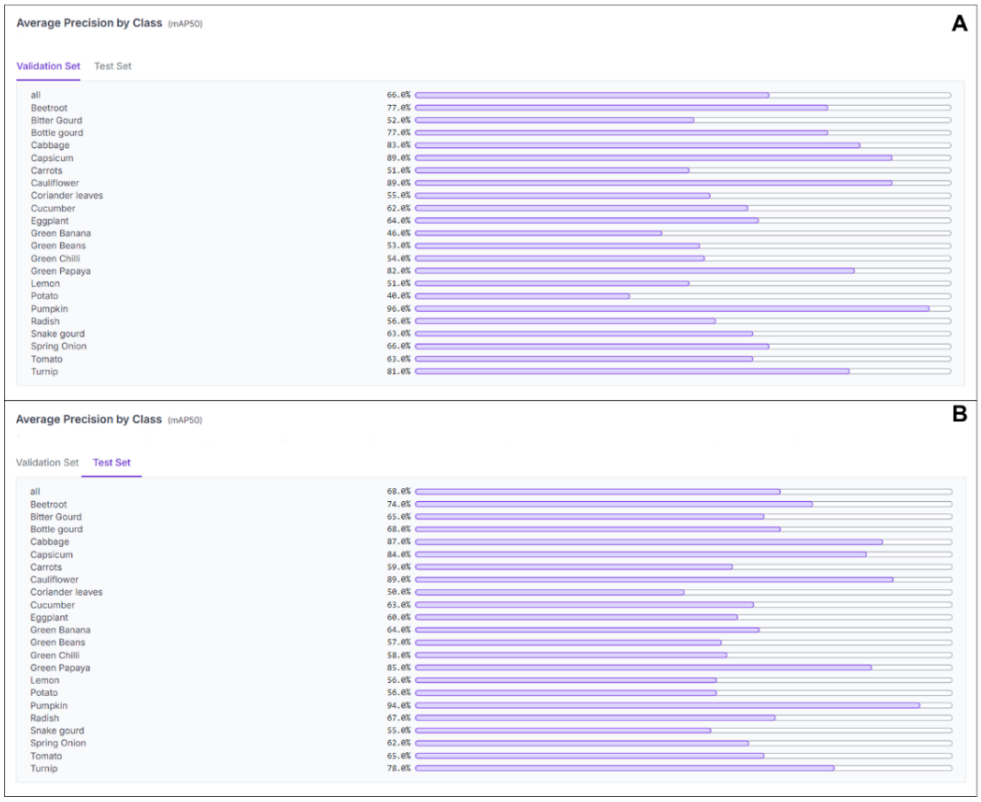

Figure 8. Class-wise average precision of the Roboflow 3.0 Object Detection (Extra Large) model on the (A) validation set and (B) test set.

YOLOv12 (Fast)

YOLOv12 Fast is optimized for real-time inference while maintaining competitive accuracy. As illustrated in Figure 9, training converges rapidly with smooth mAP and loss curves.

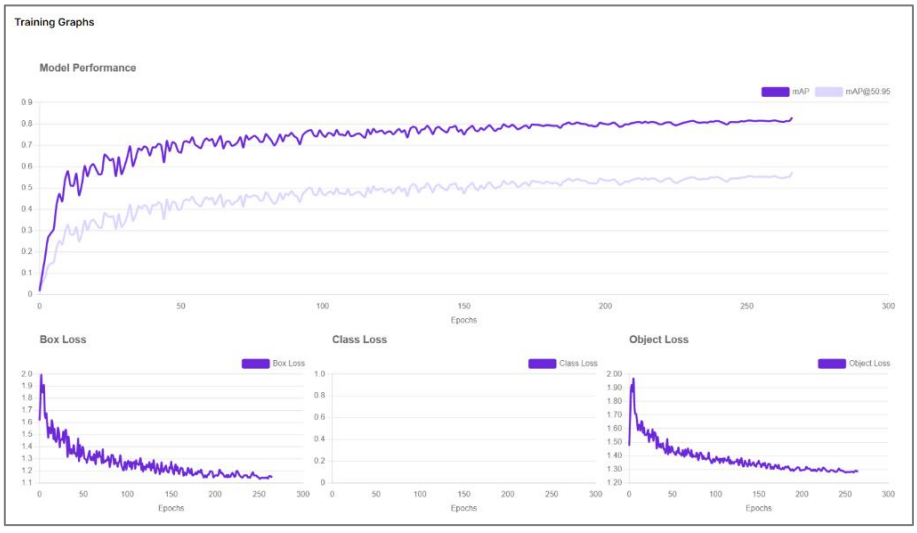

Figure 9. Training Graphs of YOLOv12 (Fast).

This model achieved an mAP@50 of 74.0% with precision of 77.4% and recall of 75.9%, significantly outperforming all earlier models. The Average Precision by class (Figure 10) demonstrates strong detection consistency across nearly all vegetable categories, validating the effectiveness of the YOLOv12 design even in lightweight configurations.

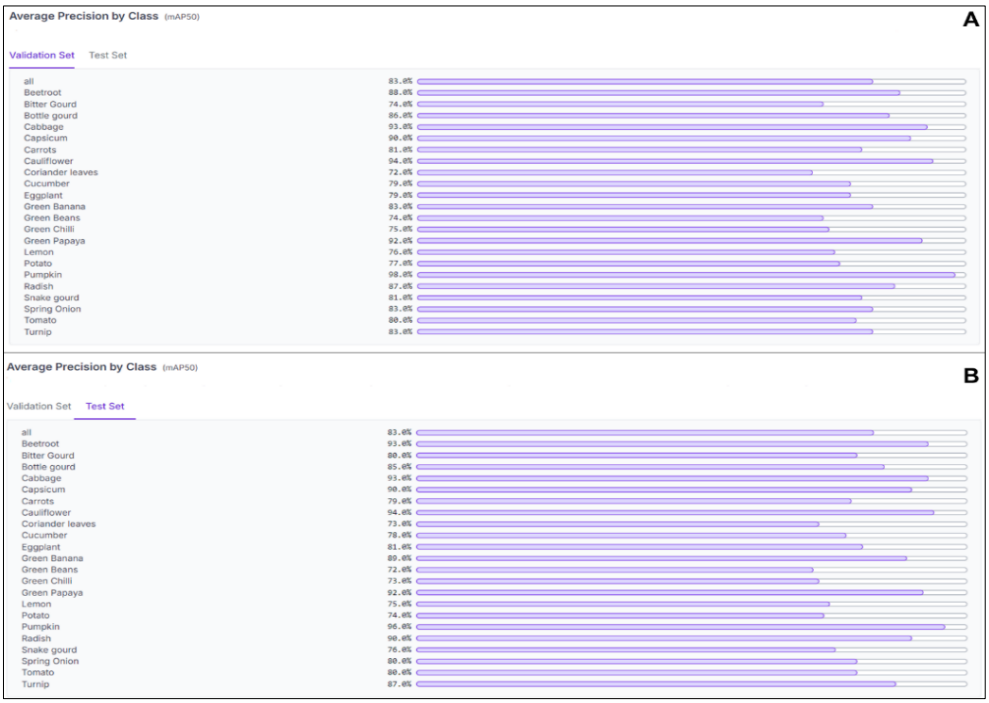

Figure 10. Class-wise average precision of the YOLOv12 (Fast) model on the (A) validation set and (B) test set.

YOLOv12 (Accurate)

The YOLOv12 Accurate variant balances speed and accuracy through enhanced feature fusion and optimization strategies. Figure 11 shows stable training with higher final mAP values than the Fast variant.

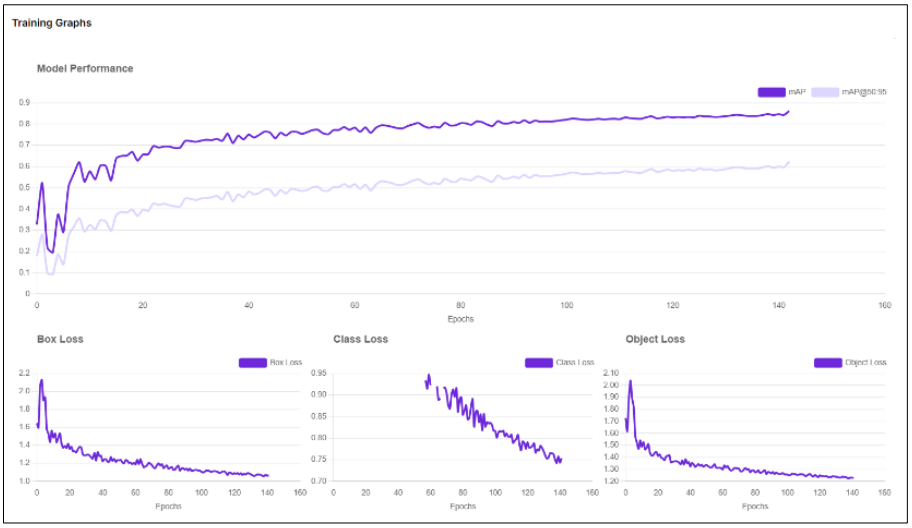

Figure 11. Training Graphs of YOLOv12 (Accurate).

This model reached an mAP@50 of 78.0%, with precision of 80.9% and recall of 79.8%. Class-wise AP values exceed 75% for most categories, indicating strong generalization (Figure 12).

Figure 12. Class-wise average precision of the YOLOv12 (Accurate) model on the (A) validation set and (B) test set.

YOLOv12 (Extra Large)

YOLOv12 Extra Large represents the highest-capacity model evaluated in this study. Training curves in Figure 13 demonstrate excellent convergence and minimal variance between training and validation metrics.

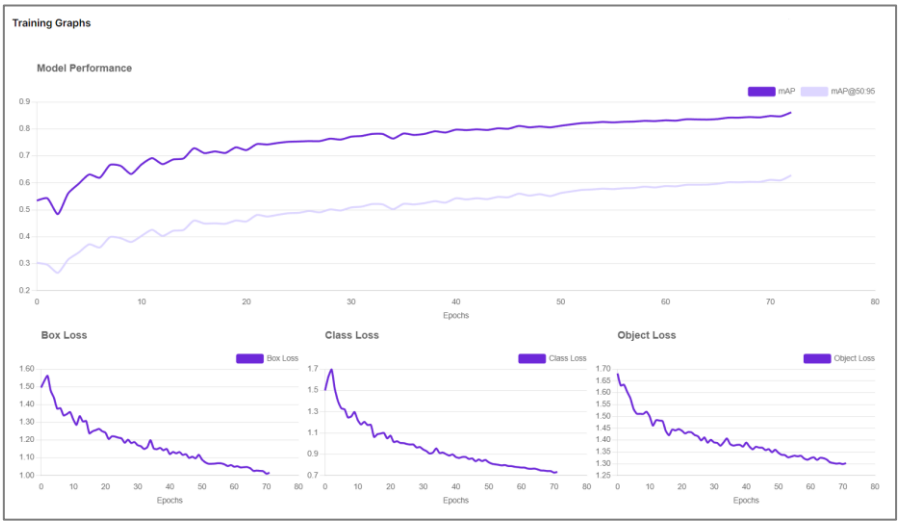

Figure 13. Training Graphs of YOLOv12 (Extra Large).

The model achieved the best overall performance, with an mAP@50 of 79.0%, precision of 81.8% and recall of 80.7%. The Average Precision by class (Figure 14) confirms superior detection across all vegetable types, including small and visu- ally similar classes. These results establish YOLOv12 Extra Large as the most effective pretrained detector for the given dataset

Figure 14. Class-wise average precision of the YOLOv12 (Extra Large) model on the (A) validation set and (B) test set.

This study comparatively evaluated the performance of several pretrained object detection models using a real-world vegetable dataset collected under natural conditions. As shown by the performance results presented in Table 1, model architecture and model capacity play a critical role in vegetable object detection performance. In particular, transformer-enhanced models and next-generation YOLO architectures demonstrate superior detection capability compared with earlier models, especially when dealing with complex backgrounds and small object instances.

Table 1. Performance of all models used in the study.

Sl No	Models	mAP@50	Precision	Recall
01	RF-DETR (Large)	67.6	67.8	67.3
02	YOLOv11 (Extra Large)	59.8	69.3	65.7
03	Roboflow 3.0 Object Detection (Accurate)	58.8	65.9	65.9
04	Roboflow 3.0 Object Detection (Extra Large)	59.4	67.5	67.3
05	YOLOv12 (Fast)	74.0	77.4	75.9
06	YOLOv12 (Accurate)	78.0	80.9	79.8
07	YOLOv12 (Extra Large)	79.0	81.8	80.7

Figure 15, which presents the performance comparison of YOLOv12 variants (Fast, Accurate and Extra Large), clearly illustrates the substantial performance improvements achieved through recent architectural advancements. To provide a deeper understanding of model behavior, the following subsection examines each trained model individually by analyzing both training dynamics and class-wise detection performance.

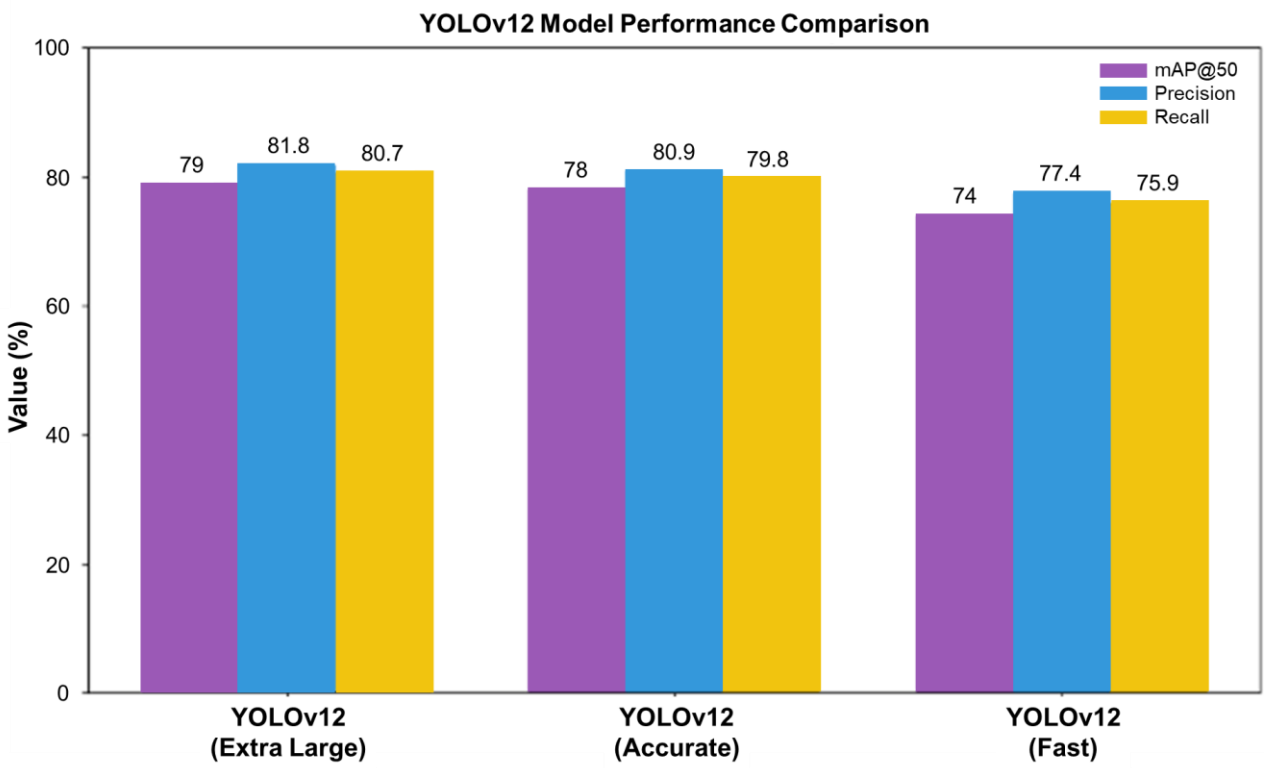

Figure 15. Comparison performance of YOLOv12 variants (Fast, Accurate and Extra Large).

YOLOv12's superior performance may be associated with improvements in feature aggregation mechanisms, architectural refinements and training optimization strategies that enhance representation capacity and detection accuracy in complex visual scenes. Recent developments in object detection architectures have shown that such design improvements can substantially improve model performance across different domains (Redmon et al., 2016; Jocher et al., 2023). However, although larger models generally provide higher detection accuracy, they also require increased computational resources, which must be considered when deploying models in real-time or resource-constrained environments (Lohumi et al., 2021; Khanam and Hussain, 2024; Kurniawan et al., 2024).

The relatively lower performance of YOLOv11, Roboflow 3.0 models and RF-DETR observed in this study suggests that recent architectural refinements may contribute more substantially to detection performance than model scaling alone under the evaluated dataset conditions. This observation is consistent with previous studies reporting that modern detection architectures can improve performance in challenging agricultural imaging scenarios characterized by variable lighting, occlusion and complex backgrounds (Carion et al., 2020; Li et al., 2022; Wang and Liu, 2024).

This study presented a comprehensive comparison of multiple pretrained object detection models using a real-world vegetable dataset from Bangladesh. Among all evaluated models, YOLOv12 Extra Large achieved the best overall performance, outperforming earlier YOLO variants, Roboflow models and RF-DETR. These results indicate that modern pretrained detectors can effectively generalize to agricultural domains when appropriate model architectures and model capacities are selected. Similar findings have been reported in previous studies demonstrating the effectiveness of transfer learning and modern object detection frameworks for agricultural vision applications (Li et al., 2022; Wang and Liu, 2024; Tapia-Mendez et al., 2025). Overall, this work provides practical insights for researchers and practitioners seeking to deploy object detection systems in agricultural environments.

Conclusions

This study presented a comparative evaluation of several modern pretrained object detection models using a real-world vegetable dataset from Bangladesh. Under identical training conditions, clear performance differences were observed among the evaluated architectures, with YOLOv12 Extra Large achieving the best overall detection accuracy across most vegetable classes, outperforming earlier YOLO variants, Roboflow models and RF-DETR. These results indicate that recent architectural improvements in object detection models can significantly enhance performance in complex agricultural imaging environments. Overall, the findings demonstrate that modern pretrained detectors can effectively generalize to agricultural datasets when appropriate model capacity and architecture are selected, providing useful guidance for the development of deep learning–based agricultural monitoring systems. Future work may explore larger and more diverse agricultural datasets, additional transformer-based detection architectures and real-time deployment scenarios in precision agriculture, as well as lightweight models optimized for edge devices used in smart farming environments.

Author Contributions

Sakibul Hasan Chowdhury performed all the experiments, data analysis, methodology and drafted the main manuscript text. Nalanda Dewan Goongoon contributed to the design and methodology of the study. Yakup Kutlu supervised the study and provided critical revisions to the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflict of Interest Statement

The authors declare that they have no competing interests.

Ethical Approval Statement

No ethics committee permissions are required for this study.

Data Availability Statement

The datasets used in this study are publicly available and can be accessed through their respective sources. All object detection experiments, including training and evaluation of YOLO-based and transformer-based models, were conducted using the Roboflow 3.0 platform.

References

Alamsyah, A.N., Prasetyo, N.A., Wibowo, F.M., Amrulloh, A., Sari, P.K. (2023). An implementation of YOLOv5 and flutter framework for fruit and vegetable object detection. *In*: The Proceeding Book of 2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Malang, Indonesia, 259-264, doi: 10.1109/COMNETSAT59769.2023.10420564

Atasoy, H., Kutlu, Y. (2022). Parameter Selection in Elliptical Fourier Series for Leaf Classification. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 24(71), 375-381. https://doi.org/10.21205/deufmd.2022247104

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. (2020). End-to-end object detection with transformers. *In*: The Proceeding Book of 16th European Conference on Computer Vision (ECCV 2020), Glasgow, UK, 213-229, doi: 10.1007/978-3-030-58452-8_13

Camgözlü, Y., Kutlu, Y. (2023). Leaf image classification based on pre-trained convolutional neural network models. Natural and Engineering Sciences, 8(3), 214-232.

Camgözlü, Y., Kutlu, Y. (2020). Analysis of Filter Size Effect in Deep Learning. Journal of Artificial Intelligence with Applications, 1(1), 20-29, doi : 10.5281/zenodo.4540543

Jahan, S., Alam, B.S., Manzur, I., Rahman, T., Hasan, M., Gani, R., Hossain, M., Shams, K., Rashid, M.R.A. (2025). Smartphone-based multi-criteria vegetable object detection dataset from Bangladesh. Data in Brief, 112281, doi: 10.1016/j.dib.2025.112281

Jocher, G., Chaurasia, A., Qiu, J. (2023). Ultralytics, YOLOv8. https://docs.ultralytics.com/models/YOLOv8/, version (01/2026).

Khanam, R., Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. arXiv preprint, arXiv:2410.17725, doi: 10.48550/arXiv.2410.17725

Khanna, S., Chattopadhyay, C., Kundu, S. (2024). Enhancing fruit and vegetable detection in unconstrained environment with a novel dataset. Scientia Horticulturae, 338, 113580, doi: 10.1016/j.scienta.2024.113580

Kurniawan, H., Arief, M. A. A., Lohumi, S., Kim, M. S., Baek, I., Cho, B.-K. (2024). Dual imaging technique for a real-time inspection system of foreign object detection in fresh-cut vegetables. Current Research in Food Science, 9, 100802, doi: 10.1016/j.crfs.2024.100802

Li, J., Qiao, Y., Liu, S., Zhang, J., Yang, Z., Wang, M. (2022). An improved YOLOv5-based vegetable disease detection method. Computers and Electronics in Agriculture, 202, 107345, doi: 10.1016/j.compag.2022.107345

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *In*: The Proceeding Book of 13th European Conference on Computer Vision, Zurich, Switzerland, 740-755, doi: 10.1007/978-3-319-10602-1_48

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). Ssd: Single shot multibox detector. *In*: The Proceeding Book of 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 21-37, doi: 10.1007/978-3-319-46448-0_2

Lohumi, S., Cho, B.-K., Hong, S. (2021). LCTF-based multispectral fluorescence imaging: System development and potential for real-time foreign object detection in fresh-cut vegetable processing. Computers and Electronics in Agriculture, 180, 105912, doi: 10.1016/j.compag.2020.105912

Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In*: The Proceeding Book of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 779-788.

Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28, 1-9.

Sakai, Y., Oda, T., Ikeda, M., Barolli, L. (2016). A vegetable category recognition system using deep neural network. *In*: The Proceeding Book of 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), Fukuoka, Japan, 189-192, doi: 10.1109/IMIS.2016.84

Tapia-Mendez, E., Hernandez-Sandoval, M., Salazar-Colores, S., Cruz-Albarran, I.A., Tovar-Arriaga, S., Morales-Hernandez, L.A. (2025). A novel deep learning approach for precision agriculture: quality detection in fruits and vegetables using object detection models. Agronomy, 15(6), 1307, doi: 10.3390/agronomy15061307

Wang, X., Liu, J. (2024). Vegetable disease detection using an improved YOLOv8 algorithm in the greenhouse plant environment. Scientific Reports, 14, 4261, doi: 10.1038/s41598-024-54540-9

Zeng, G. (2017). Fruit and vegetables classification system using image saliency and convolutional neural network. *In*: The Proceeding Book of 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 613-617, doi: 10.1109/ITOEC.2017.8122370

Zhu, C., Jia, J., Arslan, T. (2026). FVOR-YOLO: A real-time model for fruits and vegetables detection in complex supermarket self-checkout environments. IEEE Internet of Things Journal. doi: 10.1109/JIOT.2025.3643299