Monitoring the Growth of Tomatoes in Real Time with Deep Learning-based Image Segmentation

Sigit Widiyanto[1], Dheo Prasetyo Nugroho[2], Ady Daryanto[3], Moh Yunus[4], Dini Tri Wardani[5]
Department of Computer Science, Gunadarma University, Jakarta, Indonesia[1, 2]
Department of Agrotechnology, Gunadarma University, Jakarta, Indonesia[3]
Department of Information System, Gunadarma University, Jakarta, Indonesia[4, 5]

***Abstract*—Increasing agricultural productivity such as tomatoes needs to be increased, considering the consumption growth reaches 6.34% per year. Efforts to increase productivity can be made through several methods, such as counting and predicting the time of fruit to be harvested. This information is a a visual problem, so computer vision should solve it as an automation method in the industry world. With this information, the farmer can monitor the tomato fruit growth. The proposed method is a framework that has been implemented in real-time processing. To obtain growth information of tomatoes, the tomato area can be used as a region of interest (ROI) every week or another scheduled time. As the challenge of this research, this ROI can be extracted using segmentation analysis. The segmentation method used is Mask Region-Convolutional Network (R-CNN) with ResNet101 architecture. The accuracy of this method is obtained from the similarity value between the proposed method and the ground truth used, namely 97.34% using the Dice Coefficient and 94.83% using the Jaccard Coefficient. This result indicates that the method can extract the ROI information with high accuracy. So, the result can be used as a reference for the farmer to treat each tomato plant.**

***Keywords*—Deep learning; Mask R-CNN; segmentation; tomato; growth**

I. Introduction

Tomato consumption from 2016 - 2020 increased from 883.23 thousand tons to 1,084.99 thousand tons, or an increase of 6.34% from 2019 [1]. Despite an increase in production, tomato fruit has perishable properties and is classified as a climacteric fruit, where the peak of respiration and ethylene production occurs at the beginning of fruit ripening [2]. This can cause tomatoes to be easily damaged, causing a reduced supply of tomatoes at the consumer level and food insecurity. In addition, tomato production also faces future challenges in scarcity of water resources, soil salinization, and other abiotic stresses. The growth and development of tomato plants are influenced by several factors, including temperature, humidity, and altitude. If the environment does not support the growth and development of tomato plants, it will affect the productivity of tomato plants.

Increasing agricultural productivity through several methods, such as fruit counting and prediction of fruit to be harvested and early detection of environmental diseases and weeds, can be done at this time. Solutions for utilizing the latest technologies such as deep learning, the internet of things, and robotics are very effective and efficient for plant management [3]. Smart farming systems can reduce waste, increase productivity, and allow the management of more resources through remote sensing [4]. Remote monitoring through intelligent farming systems allows production yields to increase because farmers need a lot of time to solve pests, soil conditions, rain, or weeds, which can now be done through remote sensing and automation. A survey that has been conducted [5] has analyzed various articles related to deep learning technologies, and each work is compared with existing techniques. Deep learning methods have been widely used in various fields of agriculture, such as plant disease detection, crop classification, weed identification, fruit counting, land classification, obstacle detection, image translation, weather forecasting, yield prediction, and animal behavior classification [6].

Deep learning technology that has been applied to the horticultural domain for variety recognition, yield estimation, quality detection, growth, surveillance, and other detection, where this review aims to assist researchers and provide guidance to them in the use of deep learning. This guide is to understand the strengths and weaknesses that may occur when implementing deep learning. From the results of a review that has been carried out [7], there is a deep learning technique for object detection that is commonly used in horticultural crops, namely Convolutional Neural Network (CNN). There are three types of object-based detection with CNN; the first is object recognition such as LeNet, AlexNet, VGGNet, GoogLeNet, and ResNet, while the second is a combination of the first method with two-stage detection to achieve improved and accelerated detection. This method includes R-CNN, Faster R-CNN, and Mask R-CNN. As for the third type, detection is for one stage, which can immediately give the results of object boundaries according to their position. Using these models, fruit yields can be estimated automatically, and plant stress levels can be detected early. An agriculture science model with the help of deep learning techniques can be used to detect leaf diseases with images of corn, peaches, grapes, potatoes, tomatoes, and strawberries [8][9]. In this case, the image processing technique used is the CNN model for plant disease detection. The tests that have been carried out give an accuracy rate of 94.29%.

The Faster R-CNN method in the classification process of apple objects can give very accurate results [10]. A citrus fruit yield mapping system has been developed using a robotics platform [11]. The results of presenting the Fast R-CNN method used to detect citrus fruits that have been taken from different conditions (distance to fruit, camera angle, and slight

variation) with the implementation of this method give better results than the human process with an Average Precision score (AP)=0.76. The blueberry fruit calculation process has been automatically developed based on ripeness and maturity classification with deep learning methods. In this case, what is used for blueberry fruit segmentation is the R-CNN model by producing an average precision for validation and test datasets is 78.3% and 71.6% below the 0.5 intersections over union (IOU) threshold with an accuracy of 90.6% and 90.4%, respectively [12].

There is another framework, namely the YOLOv3 framework, which has been modified to YOLO-tomato, which can show better results than other advanced methods so that the YOLO-tomato method is better for detecting tomatoes in real-time with complex tomato environmental conditions [13]. The deep learning architecture used in general is a semantic segmentation architecture such as Deep Neural Network (DNN) in the field of Computer Vision (CV) [14]. The most popular DNN architectures are AlexNet, GoogleNet, VGGNet, and Resnet. Implementation of semantic segmentation techniques with DNN architecture has several problems, one of which is caused by the many parameters involved or overfitting. DNN requires high-quality labeled data and large-scale data. So, an effective solution is to build large, high-quality data sets that are difficult to achieve. Semantic detection in real-time is very important because it can be useful in autonomous systems and robotic interactions. Therefore, several new methods are adopted to increase computational efficiency, accuracy, and background noise. The semantic segmentation architecture is a fully convolutional network (FCN), ParseNet, deconvolution network, U-Net, feature pyramid network (FPN), and Mask R-CNN. The survey results show that there are many scopes of improvement in terms of accuracy, speed, complexity, and overfitting problems, so that new methods or combinations of semantic segmentation architecture are needed to increase efficiency and accuracy. In one extension of this model, [15] proposed a Mask R-CNN for object instance segmentation, which beats all previous benchmarks on many COCO challenges. This model efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. Mask R-CNN is essentially a Faster RCNN with three output branches. The first computes the bounding box coordinates, the second computes the associated classes, and the third computes the binary mask to segment the object. The Mask R-CNN loss function combines the losses of the bounding box coordinates, the predicted class, and the segmentation mask and trains jointly. In this research, Mask R-CNN is used to segment the tomato area. The segmentation is more challenging than object detection or classification because all subject areas should be identified, and the pixel that false detection would have become an error of the method. Mask R-CNN more comprehensively detecting the region rather than R-CNN with masking method.

II. METHODS

This research proposes a framework consisting of several stages, namely image data collection, image data labeling, data set separation for training and model validation, then continued by building a model and implementing segmentation using Mask R-CNN and ending with similarity testing between segmentation results using the deep learning method with manual segmentation results. In detail, the stages in the proposed framework can be seen in Fig. 1.

A. Data Collecting

All images were obtained from the primary data captured from tomatoes in a greenhouse. The data was collected from tomatoes is planted in the period from July to September 2021. The commercial tomato varieties planted and used as the data are Tora, Servo, and Tatyana. These varieties can live in the lowlands. The image was taken using a mobile phone camera and mobile robot camera with resolution 2592×1944 (5MP). The image contains multiple objects inside. The total dataset collected is about 600 images (includes 210 images Tora, 185 Servo, and 205 Tatyana). From these images, there are 2476 objects of tomato that can be identified visually. In this study, no preprocessing stage was applied to the image to be tested. Only train data needs to be labeled to support the Mask R-CNN method, and this is related to system development in actual conditions, where lighting and image transformation obtained are unavoidable. An example of an image data set can be seen in Fig. 2.

B. Image Labeling

Image segmentation aims to recognize and understand what's in the image at the pixel level. Every pixel in an image belongs to a single class, as opposed to object detection, where the bounding boxes of objects can overlap.

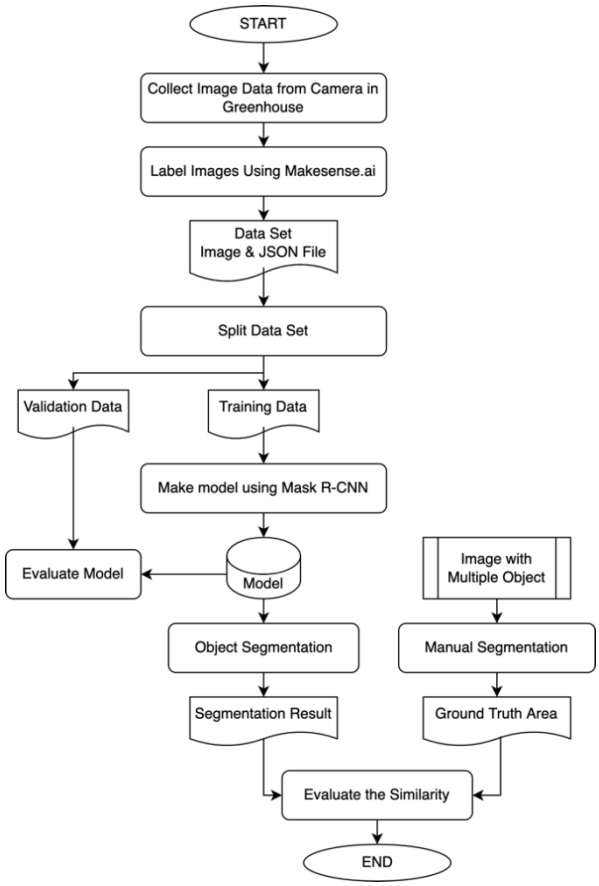


Fig. 1. Proposed Framework.

Fig. 2. Sample of Image Data Sets.

Fig. 3. Image Labeling Process using Maksense.ai.

Therefore, one preprocessing stage is needed in image segmentation, namely, image labeling. As shown in Fig. 3, an example image labeling, there are two tomatoes, so the model should be identifying all the pixels belonging to each tomato. That's where instance segmentation comes in. With instance segmentation, multiple disparate regions can belong to a single instance of a class. Each tomato in the example is annotated as an instance of the tomato class and is also given an ID such as "Tomato 1," "Tomato 2," and so forth. Therefore, the model can identify all the pixels belonging to an individual tomato even if the instance contains multiple regions.

The labeling process gives a point on the edge of the object. These points will be connected to form a polygon line. The coordinates of these points will be stored in a JSON file which must be called when executing the training model method.

C. Modelling and Testing using Mask R-CNN

Mask R-CNN is a deep learning framework that can detect objects in an image that generates a segmentation mask for each instance or commonly called instance segmentation [15]. This method runs on Faster R-CNN [16] (can be seen in Fig. 4), so that in performing mask detection, the R-CNN can be divided into three parts, namely: (i) feature extraction network, (ii) region-proposal network, and (iii) instance detection and segmentation networks.

1) Feature extraction: Mask R-CNN applies multiple backbone architecture. Some of the backbones used for Mask R-CNN are ResNet, ResNet, and FPN. In the Region Proposal Network (RPN) process, a Region of Interest (RoI) will be generated through an alignment process which will then be input for the instance detection and segmentation networks.

Mask R-CNN uses a combination of ResNet101 architecture and FPN (Feature Pyramid Network) to generate RoI features when feature extraction is performed. FPN is a basic component in the recognition system to detect objects at different scales using the same image.

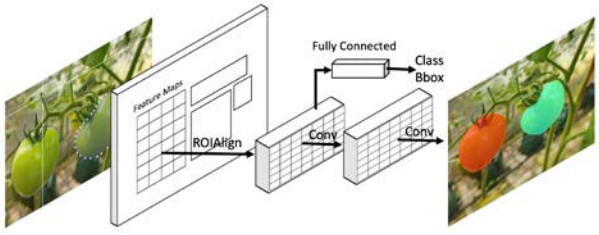


Fig. 4. Mask R-CNN Architecture.

FPN uses a variety of feature maps to produce higher-quality information. FPN is a feature extractor designed using the pyramid concept but is superior in speed and accuracy. FPN processes information in two ways, namely bottom-up (bottom-up) and top-down (top-down).

Bottom-up data processing extracts features using ResNet, in which the spatial dimension of each layer decreases while the semantic value increases. Top-down processing increases the resolution of the semantic layer, but the location of objects is not precise. FPN adds lateral connections between reconstructed layers and a corresponding feature map to help detectors better predict locations. Lateral connection is the convolution and addition operation between the two corresponding levels of the two pathways. FPN surpasses the single ConvNet because it retains semantic features at various resolutions.

From ResNet101 four feature maps were extracted (layer-1, layer-2, layer-3 and layer-4). An approach called the top-bottom pathway is used to produce the final feature map. The top-bottom pathway approach starts from the top feature map (x/32, y/32, 256) and starts down to a larger feature map by applying the up-sample operation. A 1 x 1 convolution was performed to reduce the number of channels to 256 before sampling and then added elements to the up-sample output from the previous iteration. All up-sample outputs were applied to a 3 x 3 convolution layer to produce the last four feature maps, while the fifth feature map was generated from the max-pooling operation.

2) Region proposal network: Each feature map generated in the feature extraction process will go through a 3 x 3 convolution layer. But before that, the feature map is scanned using anchor boxes with various scales and ratios. The resulting output is then forwarded to two branches, one relating to the objectivity or confidence score and the other to the bounding box regressor. A confidence score can be obtained by calculating the IoU (intersection over union) between the bounding box and the ground truth. IoU is obtained by dividing the overlapping area by the combined area of the ground truth and bounding box. The confidence score ranges from 0 to 1. The greater the Confidence score means, the higher the system confidence that the object contained in the bounding box is the object to be detected.

3) Instance detection and semantic segmentation: The segmentation instance process is carried out using a fully connected network that takes RoI as input to detect the presence of objects, bounding boxes, class labels, and confidence values. A fully Convolutional Network (FCN) is

used to perform semantic segmentation on the image by predicting the semantic class of each pixel in the bounding box. This causes each instance to display a different color according to its bounding box.

D. Manual Segmentation

A manual segmentation process is needed to generate Ground Truth data. This data is used to compare whether the segmentation process using the Mask R-CNN method is successful in approaching expert judgment. Manual segmentation is carried out using an image processing application to remove or change the background with black color or a value of 0 (see Fig. 5.b) and then convert to a binary image (see Fig. 5.c).

E. Similarity Evaluation using Dice and Jaccard Coefficient

Dice and Jaccard coefficient have been utilized to measure the segmentation accuracy. Rockefeller used these coefficients in the segmentation in cucumber seed [17]. The Jaccard coefficient measures similarity between sample sets. It is defined as the size of the intersection divided by the size of the union of the sample sets. Sample sets are obtained by mask (black, white) image on every region from segmentation result and ground truth image, to be detailed see Equation 1. Samples sets contain, i.e., the region of interest from algorithm segmentation result and manual segmentation by an expert judgment as ground truth data. The Jaccard coefficient also called Intersection over Union (IoU), is used in the Mask R-CNN evaluation model.

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (1)$$

Where X is the set of every region from segmentation and Y is the set of every region from the ground truth image, which has a true value (1 in binary image).

Then, almost like Jaccard, Dice distance is also utilized to measure the similarity, but it has different properties, Equation 2. Dice are used as an optimist measurement which gives the double weighting value on the true positive value (intersection of two regions).

$$D(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|} \quad (2)$$

Where 2 |X ∩ Y | is the double weighting of number value which is intersected on two data sets. While |X| and |Y | are the number of data sets which are having true value (1).

(a) (b) (c)

Fig. 5. Manual Segmentation Process. (a) Original Image, (b) Remove or Change Backgorund, (c) Convert to Binary Image.

III. Results

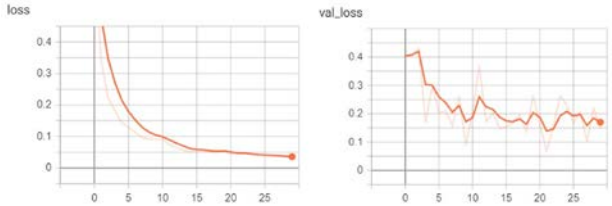
A. Model Evaluation

To conduct training and validation, 530 data sets were used consisting of various types of age and tomatoes. Of all the data sets, 70% are used as training data sets and 30% for testing data sets. The evaluation of the results of the model form can be seen in Fig. 6. Fig. 6.a shows the training loss, which goes towards 0 as the epoch value increases, while Fig. 6.b shows the validation loss, which is 0, although it is not as good as the training loss. However, these two graphs are convergent, showing that the model formed is fit, meaning that the model can predict the data set outside the training data.

B. Segmentation Result

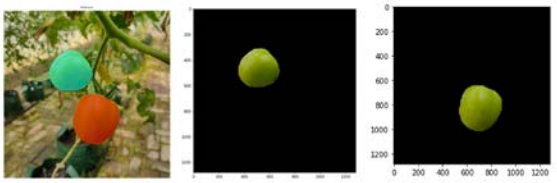
After the model is formed and thorough evaluation, it is declared fit. Then the model is then used to perform image testing segmentation. Mask R-CNN can perform segmentation by providing a mask, and at the same time assigning class labels as class numbering "Tomato-1" and "Tomato-2" so that objects can be separated. In addition, this method can also provide a bounding box accompanied by a confident value from the results of the class classification. In Fig. 7.a can be seen the segmentation results, where the confident value can also be seen. Fig. 7.b and Fig. 7.c are the results of extracting fruit areas "Tomato-1" and "Tomato-2".

The next process is to calculate the number of pixels from the area of each tomato. The first process is to convert the image into a binary image or a value of 0 (black) and 1 (white). Furthermore, the total number 1 in the image can show the number of pixels of each object. Fig. 8 shows the result of converting an image into a binary image.



(a) (b)

Fig. 6. Model Evaluation. (a) Training Loss, (b) Validation Loss.



(a) (b) (c)

Fig. 7. Segmentation Result. (a) Image Segmented with Mask, (b) Tomato-1 Area, (c) Tomato-2 Area.

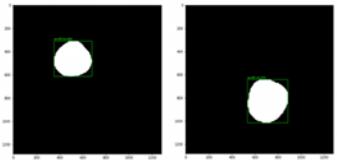


Fig. 8. The Result of Binary Conversion Process.

C. Segmentation Evaluation

Segmentation results are evaluated using the Dice and Jaccard coefficient. This method is used to assess the similarity between the proposed method result and the ground truth obtained through manual segmentation. In Fig. 9, the segmentation results using the proposed method do not have smooth or imperfect edges like manual segmentation results. Thus, this deviation makes the difference in the number of pixels between the two. The evaluation results show that the first image has a similarity value based on the Dice coefficient of 97.90. This value shows 97.90% of the pixels in the segmentation area with the proposed method are the same as the manual segmentation results, likewise, with the Jaccard coefficient, which shows a value of 95.90. Jaccard always gives less value than Dice.

In Fig. 10, the graph of the similarity evaluation results on 30 data sets used as testing data can be seen. The graph shows that the average evaluation result is above 90%. Only one image can be segmented with a similarity value of about 87%.

The average evaluation results for 30 images with the Dice coefficient is 97.34% and using the Jaccard Coefficient is 94.83%. These results indicate that the Mask R-CNN can segment well on the image of tomatoes from various types of tomatoes and tomato colors. Thus, the calculation results in the form of some pixels from the tomato area can be declared valid so that the research can be continued with analyzing the development of tomatoes.

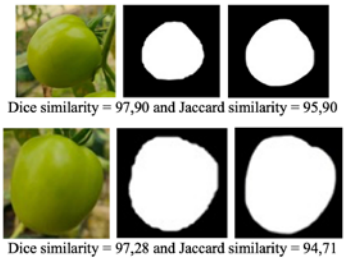


Fig. 9. Segmentation Evaluation. Left: Original Image, Center: Proposed Method Result, Right: Manual Segmentation Result.

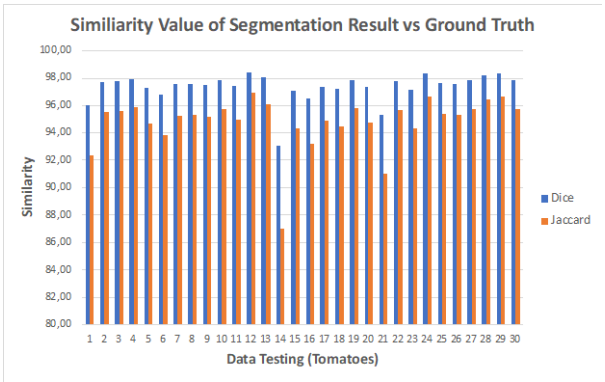


Fig. 10. Similarity Value of Segmentation Result vs Ground Truth.

IV. Discussion

One of the indicators related to the increase in tomato production is the growth of tomato fruit after flowering. The camera can capture fruits formed for a week as objects that have characteristics, namely round or oval or other shapes according to the type of plant. Tomato fruit growth can be seen from the area of the fruit. Although the picture of the fruit must be taken from a consistent direction, for example, in the first week after fruit formation is taken using a camera from an angle of 180 degree or straight from the top of the tree, then in the following week, it is taken from above as well. Likewise, if taken from the side of 90 degree.

In Fig. 11, you can see a graph of the increase in the ten observed fruits (not all of them are shown due to the clarity of the graph). Some fruits have different growth patterns. Knowing each fruit's growth pattern or average can be recommended regarding environmental engineering and proper fertigation. Thus, fruit growth can be evenly distributed, and harvest targets can be achieved.

Tomato fruit development during four weeks of the observation showed good development and was detected by increasing the number of pixels area. Tomato fruit develops every week marked by increasing fruit dimensions and fruit color that changes from light green to a red tinge and finally red. The time required by tomatoes from flowering to harvesting was 5-6 weeks [18]. In Fig. 12 can be seen the average of tomato fruit growth in a week, which shows how each tomato grows and require special handling for such fruit like tomato 5 and 6. According to [19], the weight of small vegetable tomatoes has an average weight of < 50 g per fruit, medium size with a weight range per fruit of 50 - 70 g, and large size with a weight per fruit of > 70 g. This size unit must be converted in some pixels by volume calibration, so the unit can match the image processing standard.

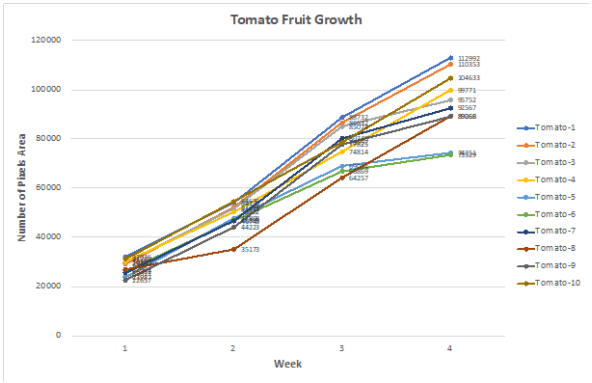


Fig. 11. Tomato Fruit Growth in 4 weeks after Fruit Formation

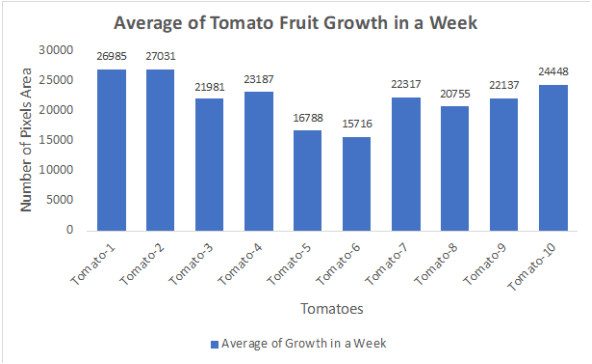


Fig. 12. Tomato Fruit Growth in 4 weeks after Fruit Formation.

V. Conclusion and Future Work

The Mask R-CNN method can perform properly and has accuracy as indicated by the similarity value measured by the Dice Coefficient and Jaccard Coefficient with an average value of 97.34% and 94.83%, respectively. This similarity value indicates that this method can be used to find the Region of Interest area of tomato objects so that it can be used to measure tomato growth. The entire system has worked in real-time with various problems such as lighting, morphology, and transforming fruit shapes and other objects. The real-time segmentation method is difficult to implement using K-Means, SVM, or Neural Network methods. For future work, it is necessary to calibrate the fruit weight unit (kg) into the volume, while in terms of the proposed method, a volume estimation process must be added. Thus, higher accuracy will be obtained regarding the analysis of tomato fruit growth based on the image.

Acknowledgment

This research was funded through the Higher Education Applied Research grant program under the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia with contract numbers 309/E4.1/AK.04.PT/2021 and 09.17/LP/UG/VII/2021.

References

[1] Food and Agriculture Organization of United Nations, Summary Report, Joint Fao/Who Meeting on Pesticide Residues, October 2021.

[2] Chen Y, et.al., “Ethylene receptors and related proteins in climacteric and non-climacteric fruits,” Plant Sci, vol. 276, pp. 63–72, 2018.

[3] Kavitha, “ Deep Learning for Smart Agriculture,” Int. Journal of Engineering Research & Technology, vol. 9, no. 5, 2021.

[4] M. Vengateshwaran, N. Sumithra, S. P. Rani, and B. Pravalika, “A Deep Learner based Smart Precision Agriculture System using Machine Learning Techniques,” Int. Journal of Engineering Research & Technology, vol. 9, no. 5, 2021.

[5] C. Ren, D. K. Kim, and D. Jeong, “A Survey of Deep Learning in Agriculture: Techniques and Their Applications,” Journal of Information Processing Systems J Inf Process Syst, vol. 16, no. 5, pp.1015-1033, October 2020.

[6] N. Zhu, et.al., “Deep learning for smart agriculture: Concepts, tools, applications, and opportunities,” Int. Journal of Agricultural and Biological Engineering, vol. 11, no. 4, pp. 21-28, 2018.

[7] B. Yang and Y. Xu, “Applications of deep-learning approaches in horticultural research: a review,” Horticulture Research, vol. 8, no.123, 2018.

[8] Md. Tariqul Islam, “Plant Disease Detection using CNN Model and Image Processing” in in International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 10, October 2020.

[9] S. Widiyanto, R. Fitrianto, and D. T. Wardani, “Implementation of Convolutional Neural Network Method for Classification of Diseases in Tomato Leaves,” Fourth International Conference on Informatics and Computing (ICIC), pp. 1-5, 2019.

[10] H. Shaikh, Y. Wagh, S. Shinde, and S. M. Patil, “ Classification of Affected Fruits using Machine Learning,” Int. Journal of Engineering Research & Technology (IJERT), vol. 9, no. 3, 2020.

[11] J. P. M. Galdames, C. E. Milhor, and M. Becker, “Citrus fruit detection using Faster R-CNN algorithm under real outdoor conditions,” Proceedings of the 14th International Conference on Precision Agriculture, June 24 – June 27 2018.

[12] X. Ni, C. Li, H. Jiang, and F. Takeda, “Deep learning image segmentation and extraction of blueberry fruit traits associated with harvestability and yield,” Horticulture Research, vol. 7, no. 110, 2020.

[13] M. O. Lawal, “Tomato detection based on modified YOLOv3 framework,” Scientific Reports, vol. 11. no. 1, 2021.

[14] Tapasvi, N. U. Kumar, and E. Gnanamanoharan, “A Survey on Semantic Segmentation using Deep Learning Techniques,” Int. Journal of Engineering Research & Technology (IJERT) vol. 9, no. 5, 2021.

[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” Proceedings of the IEEE international conference on computer vision, pp. 2961–2969, 2017.

[16] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” Advances in neural information processing systems, pp. 91–99, 2015.

[17] R. T. Rockafellar, Variational Analysis, Springer-Verlag, 2005, Number ISBN 3-540-62772-3.

[18] A. Daryanto, M. R. A. Istiqlal, Kalsum, and R. Kurniasih, “Penampilan karakter hortikultura beberapa varietas tomat hibrida di rumah kaca dataran rendah,” Journal Agron Indonesia, vol. 48, no. 2, pp. 157–164, 2020.

[19] M. Syukur M, E. Helfi, and R. Hermanto, Bertanam Tomat di Musim Hujan, Jakarta (ID): Penebar Swadaya, 2015.