International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 2
ISSN: 2088-5334

Increasing Precision of Water Sprout Detection based on Mask R-CNN with Data Augmentation

Intan Sari Areni [a,*], Nurul Maulidyah [a], Indrabayu [b], Anugrayani Bustamin [b], Azran Budi Arief [a]

[a] Department of Electrical Engineering, Hasanuddin University, Makassar, 90245, Indonesia
[b] Department of Informatics Engineering, Hasanuddin University, Makassar, 90245, Indonesia
Corresponding author: *intan@unhas.ac.id

***Abstract*—This study evaluated the detection performance of four Mask R-CNN models trained in different scenarios. The first two scenarios are trained with a learning rate of 0.01 using data augmentation on the training data. The other two scenarios are trained with a learning rate of 0.001 and the same as previously, using augmentation on training data. These models are trained to detect water sprouts in cacao plants. The original data used are obtained from photographed pictures on the cocoa farm. As much as 150 images, the data is divided into 120 images for training data and 30 images for testing data. In previous studies, the model was trained without performing data augmentation, so that the amount of data trained was less than this study. Data augmentation is implemented to compromise the small amount of data and prevent over-fitting during the model training process. This process uses six augmentation parameters, namely horizontal flip, blur using Gaussian blur, contrast modification using linear contrast, color saturation alteration, cropping the sides of the image randomly by 50 pixels, and rotating the image. The test is carried out by varying the threshold value in the range of 0.6 to 0.9. The results obtained indicate that the model trained with a learning rate of 0.001 with data augmentation can detect objects better than other models with an F1score of 0.966 at a threshold of 0.8. This research will be developed to create a water sprout cutting robot in the future.**

***Keywords*—Image classification; object detection; feature extraction; mask R-CNN; data augmentation.**

Manuscript received 13 Oct. 2021; revised 15 Dec. 2022; accepted 5 Mar. 2023. Date of publication 30 Apr. 2023.

I. INTRODUCTION

Water sprouts are part of the plant that farmers often consider parasites. They will overgrow along with the cocoa tree's growth [1], [2]. A non-directed growth will produce a plant canopy that generally grows lengthwise upwards with a single trunk or branch. The strong dominance of the apical (shoot tip) at the tip of the plant spurs the plant to continue growing upward. In cacao plants, water sprouts can also cause Cherelle wilt due to competition between water sprouts and fruit buds in obtaining photosynthetic results [3], [4], which has resulted in a decrease in cocoa production. Cocoa is one of the leading commodities originating from South America and has been developed in Indonesia since 1930 [5]. As the third-largest cocoa producer globally, Indonesia faces a crisis where there has been a decline in cocoa production in the last ten years. Increasing production can be done by applying automation technology in agriculture and, simultaneously, reducing dependence on the availability of human labor [6], [7].

Research on automation in agriculture has been carried out over the last few decades [8]–[11]. However, its implementation has many obstacles, such as the complexity of field operations and inconsistencies in cropping systems that hinder the implementation of automation technology in plantations [12], [13]. Due to the large area of the cocoa plantation, a system is needed to detect water sprouts in each existing cocoa plant to simplify and speed up the pruning process for each water sprout grown on the cocoa tree.

Water sprouts can be detected and classified based on the color of the leaves, ranging from green, pink, brown, and several other color variations [2]. Many algorithms can be used to detect leaf objects in the image. In Zhang et al. [14] research, the individual leaf was detected by examining leaf veins to estimate its location and direction using the SKEDET method. Another study detected the disease in rubber tree leaves using edge detection techniques, namely Sobel edge detection [15]. Furthermore, leaf angle distribution measurements were performed by detecting individual leaves using the SfM-PCNN (structure from motion-pyramid CNN) method, in which leaf borders are drawn while minimizing the

influence of the inner leaf texture [16]. Chen et al. proposed a shape-based leaf segmentation method that performs leaf segmentation using a continuous function and produces precise leaf edge contours [17]. In another study, Wang et al. detected overlapping leaves using the Sobel operator and the Chan-Vese algorithm [18]. Apart from the methods mentioned earlier, deep learning methods are also widely used by researchers to detect objects.

The deep learning method is widely used in various applications such as text classification, speech recognition, and image recognition. This method uses non-linear transformation functions arranged in layers where the resulting learning model can better represent the training data. It also aligns with the growing ability of processors and graphics processors to process large amounts of data [19].

Deep learning requires large amounts of data for the training process, and more data generally results in better model performance [19], [20]. However, collecting data on water sprouts with various variations is challenging. Therefore, with a limited amount of data, increasing the amount without diminishing the data format's integrity is essential to disappear the shortage of data sets.

Song et al. apply random crop and rotate techniques to augment data. Besides, data enhancement was also carried out and then trained using the Inception3 and MobileNet models. The trained model is then embedded in an Android device to classify endangered animal species. The best accuracy result is 89% by comparing the two models obtained using the MobileNet model [21].

In Kutlugün, Sirin, and Karakaya's research, face recognition was conducted using the CNN model. For the training process, augmentation was implemented by applying several different filters to the data set to increase data variety during the training process. Furthermore, the resulting model is analyzed to determine which filter provides the most effective classification result [19].

Furthermore, Park, Lee, and Park performed data augmentation on human body parts for training in human pose estimation models. This method offers to crop the body part in the image to be trained so that no image segment is too small, and there is no redundancy in the augmented data [22].

Almutairi and Almashan researched instance segmentation on newspaper elements (page headers, articles, and advertisements). A horizontal flip technique on 50% of the data set images is carried out to increase data sets. Thus, the training process focuses on the layout of the article instead of the text contained in it. The model used to train the dataset is the Mask R-CNN [23]. Islam *et al*. also carried out research related to CNN by analyzing the augmentation process. Some of the techniques used are re-scaling, zooming, shearing, rotation, width, and height shifting. The resulting model with augmented data achieves an accuracy of 97.12%. This performance was 4% higher when compared to the model without augmentation [24].

In this study, the Mask R-CNN method was applied to detect water sprouts in cocoa plants, and data augmentation was applied to develop previous studies [25] to improve the detection results of water sprouts. The Mask R-CNN is a deep learning model that uses a regional convolutional neural network (R-CNN) for object detection, classification, and segmentation [26], [27]. The training data augmentation technique is performed to add variation. First, the data is labeled on each water sprout, then trained using Mask R-CNN. The training process is carried out four times by applying different parameters in each training process, as shown in Figure 1. Thus, the performance of each training model with and without data augmentation at 0.01 and 0.001 learning rates can be obtained.

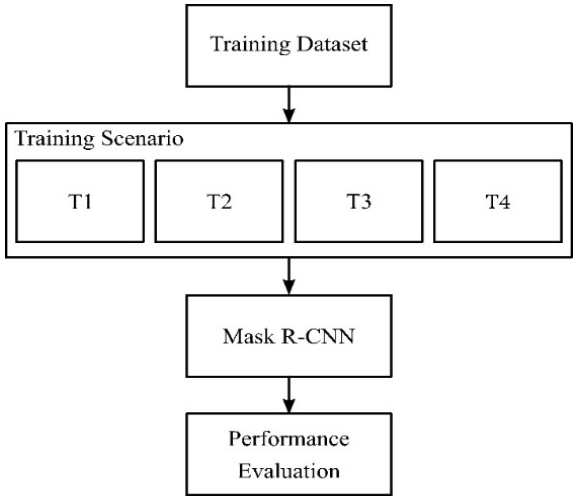

Fig. 1 Process block diagram

II. Materials and Method

A. Mask R-CNN

Instance segmentation detects objects and segments objects simultaneously. The Mask R-CNN algorithm can detect water sprouts by implementing an instance segmentation process that produces three outputs: class, bounding box, and mask. The visualization results from detecting water sprouts will make obtaining the water sprouts' position on each existing cocoa tree easier. The Mask R-CNN algorithm is an algorithm that implements the in-depth learning process in AI (Artificial Intelligence). Mask R-CNN is the development of Faster R-CNN [28], which implements semantic segmentation using the FCN (Fully Convolutional Network) algorithm [29]. Developments of Faster R-CNN include ROI Align introduced as a substitute for RoI Pooling in Faster R-CNN [30]. Since RoI Pooling is not aligned based on the top pixel one by one (Pixel-to-pixel alignment), this has no significant impact on the bounding box but significantly impacts mask accuracy. The mask accuracy after using RoI Align significantly increased from 10% to 50% [31].

Semantic segmentation is introduced to actualize the separation of the relationship between mask and class prediction, where the mask branch only performs semantic segmentation, and class prediction assignments are assigned to other branches. Thus, it differs from the original FCN network, which original FCN also predicts the type of mask it has when predicting masks. The structure of the Mask R-CNN network is shown in Figure 2.

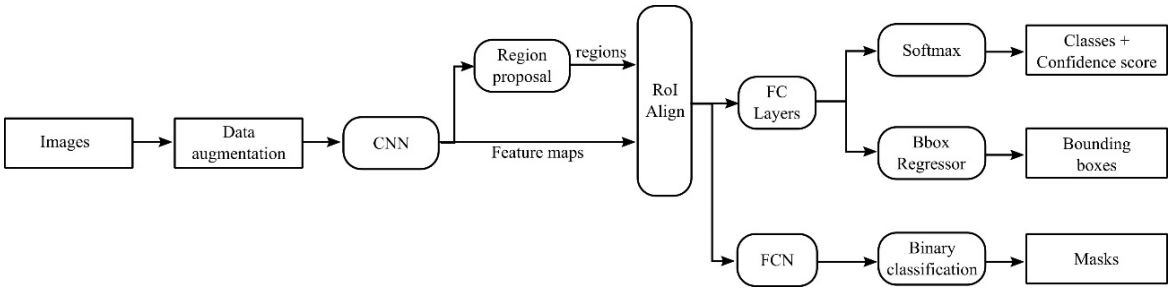

Fig. 2 Mask R-CNN architecture

The Mask R-CNN passed through a training and testing process. The training process is carried out in several stages: data preprocessing, data augmentation, feature extraction, region network proposals, RoI Classifier & bounding box regressors, and segmentation masks. In preprocessing, the image on the training dataset is labeled for each object to be studied for its features using a polygon shape; thus, the resulting bounding box resembles its shape. The next stage is data augmentation, where the artificial addition of data is carried out to obtain diverse data.

The feature extraction process used in this system is a region-based convolutional neural network (CNN). Ross Girshick introduced this method to solve the number of regions selected using the selective search method to extract 2000 regions from the image. These regions are then entered into a convolution neural network for feature extraction [32]. However, the training process is prolonged. Girshick created the Fast R-CNN method to optimize the training process time in which the region entered in the image is not extracted using selective search. However, the image is directly inserted into the convolution neural network to produce feature maps. The resulting feature maps extract the proposal region using a selective search algorithm. The Mask R-CNN is the Faster R-CNN method, which extracts the proposal region using RPN [33]. The difference between the Mask R-CNN and the previous method is the application of RoI Align and adding branches to carry out instance segmentation. For Masks R-CNN, several backbones can be used, namely ResNet50, ResNet101, ResNet152[34], [35], which will be combined with the feature pyramid network (FPN). ResNet will detect low-level features (edges and corners) in the initial layer, and the next layer will detect higher-level features, respectively, as shown in Figure 3. Whereas the feature pyramid network enhances the standard feature extraction pyramid by adding a second pyramid, which takes high-level features from the first pyramid and passes them to the bottom layer. The pyramid network feature allows each level to access lower and higher-level features to avoid missed segmentation of objects in the next stage [31], [36].

The network proposal will scan and display the ground truth bounding box and the predicted bounding box at the region stage. The ground truth bounding box is a bounding box obtained from the image's annotation process. In contrast, the predicted bounding box is obtained from predictions of areas containing object areas through the feature extraction process. At this stage, a refinement process is also carried out on the bounding box (ground truth and prediction) so that the bounding boxes do not overlap.

In the next stage, the region of interest generated from the proposed network region will classify objects using the highest Intersection over Union value. Intersection over Union is the slice value between the ground truth bounding box and the predicted bounding box. Higher Intersection over Union value results in a smaller slice between the ground truth bounding box, the predicted bounding box, and the area classified as a class in the detection process. At this stage, the best bounding box will also be given based on the Intersection over Union (IoU) value for the object contained in the image.

The segmentation mask stage is a process that takes the positive RoI that has been obtained from the RoI Classifier & bounding box regressor stage. The segmentation masks stage provides segmentation of objects that have been classified using the Fully Convolutional Network (FCN) algorithm, which plays a role in Semantic Segmentation, to provide a mask for each pixel per 3 pixels of the object. FCN performs a dense prediction that will differentiate each object's pixels in the image and display each object pixel with a different label/color.

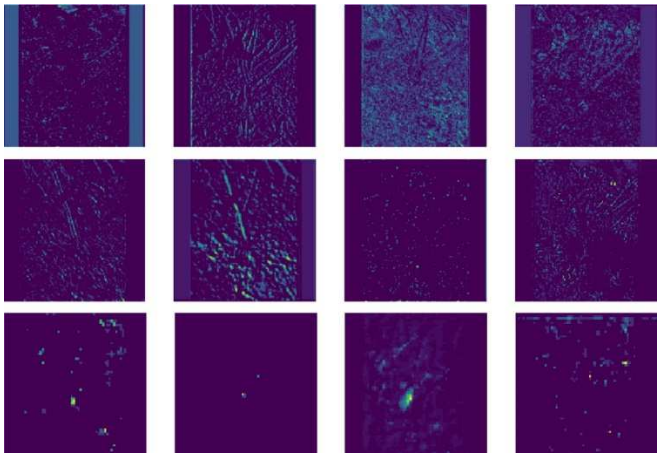

Fig. 3 The example of feature maps produced by ResNet101

B. Dataset

The training and testing data are collected from cocoa farms located in Wajo Regency, South Sulawesi, Indonesia. Data were taken using a Nikon Coolpix P610 camera with a 100-200 cm distance and a height of 100 cm, as shown in Figure 4. This selection of distance and height, considering that water sprouts on the main stem of the cocoa plant and the spacing between trees are 200-400 cm, which is the standard spacing in cocoa cultivation practice. From this data collection, 150 image samples measuring 3120 × 4160 pixels were obtained and stored in the JPEG format. The sample data was then divided with a ratio of 80:20 for the training data.

1) Preprocessing data

The data collection is 150 images with 120 training data and 30 testing data. In preprocessing, image data is labeled manually using VIA (VGG Image Annotation) to obtain annotation values in the form of object coordinates in the image used in the training stage, a ground-truth bounding box. The resizing and padding processes are also carried out to uniform the image size to 1024 × 1024 pixels at the data preprocessing stage, as shown in Figure 5.

2) Data augmentation

The training data used in this study were 120 image data. The number of 120 images is very minimal for the training process using the Mask R-CNN algorithm. This data augmentation process is used to avoid over-fitting [24]. Overfitting occurs when the accuracy of the training data is higher than the testing data. Thus, data augmentation is implemented to prevent over-fitting during the model training process. The data augmentation process makes it possible to create new data based on the data collected as input data, as shown in Figure 6. This data augmentation process uses six parameters: horizontal flip, blur using Gaussian blur, contrast modification using linear contrast, modifying color saturation, cropping the sides of the image randomly by 50 pixels, and rotating the image.

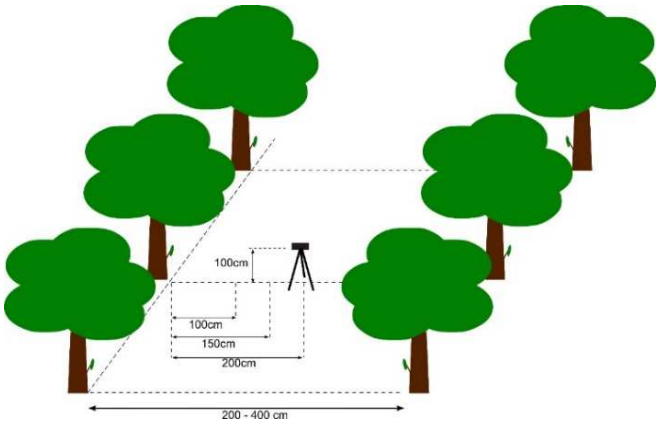

Fig. 4 Data collection scenario

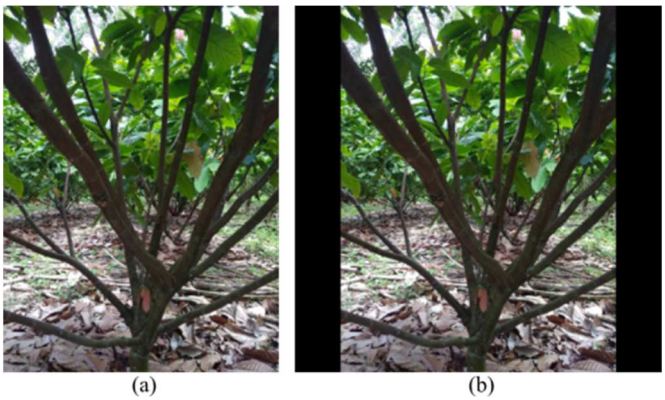

Fig. 5 Examples of training data (a) original data with 3120×4160 size, (b) resized image to 1024×1024

C. Training the Model

In the training process, the Mask R-CNN is used to study water sprout features with ResNet101 as the backbone and COCO dataset as the pre-trained model [37]. COCO is a vast dataset with 328k image data divided into 91 class categories. It is often used for object detection and image segmentation with the pre-trained model used to transfer learning from large datasets so the model used during training can learn the features of the data more quickly.

The training process was carried out with four scenarios with 50 epochs each to produce four different models to compare the performance. The training combines learning rates (LR) of 0.01 and 0.001 with data augmentation to determine the optimal configuration among the four scenarios. Details of the training scenario are as follows:

- T1: Training using LR 0.01 without data augmentation
- T2: Training using LR 0.01 with data augmentation
- T3: Training using LR 0.001 without data augmentation
- T4: Training using LR 0.001 with data augmentation

D. Detection Model

The Mask R-CNN method proceeds through three prediction steps to produce object detection. The first is the Region Proposal Network to determine the Region of Interest that contains objects to become proposals (see Figure 7(a)) based on a trained model.

The next step is to classify the proposal using the heads classifier to generate bounding boxes and confidence scores and determine the object class, as shown in Figure 7(b). The last step is to apply masking to the object detected in the previous stage, as shown in Figure 8, to extract objects from the background at the pixel level. The extraction process by FCN classifies each pixel of the water sprout object contained in the bounding box.

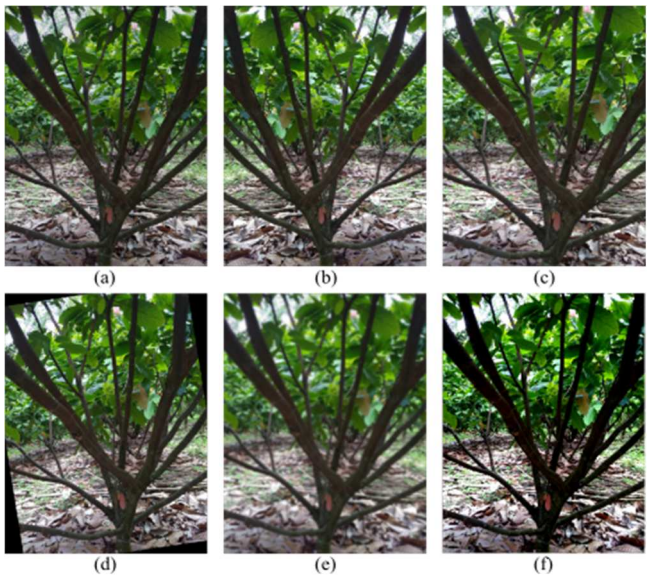

Fig. 6 Examples of data augmentation (a) multiply, (b) horizontal flip, (c) cropping, (d) rotate, (e) blur, (f) contrast

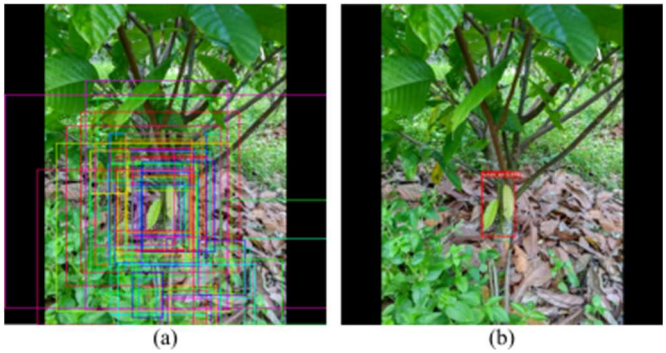

Fig. 7 Detection process (a) proposal result from RPN process, (b) classified proposal result example

E. Evaluation Metric

The recall, precision, and F1-score evaluate learning model performance. Recall calculates the ratio of water sprouts to the test data detected by the system (equation (1)). The precision score calculates the positive water sprout detection ratio to the system's overall detection of water sprouts (equation (2)). Because recall and precision calculate two different ratios, it is necessary to calculate the F1-score to determine the harmonic mean of metric recall and precision, as in equation (3). This metric is also used due to the non-symmetrical number of false-negative and false-positive data. F1-score measurements produce values in the range of 0-1, where the greater value indicates a better performance.

$$recall = \frac{TP}{TP+FN} \quad (1)$$

$$precision = \frac{TP}{TP+FP} \quad (2)$$

$$F1score = \frac{2 \cdot precision \cdot recall}{precision+recall} \quad (3)$$

TP, FP, and *FN* are obtained from confusion matrix calculations, where *TP* is True Positive detection, *FP* is False Positive detection, and *FN* is False Negative detection.

III. Results and Discussion

Experiments were carried out with four training scenarios and by changing the threshold values. The optimal F1-score of the previous study was obtained at a threshold of 0.6 [19]. So, in this study, the threshold range used was between 0.6 and 0.9 to evaluate the detection system's performance increase. In the first scenario, training (*T1*) is conducted with LR 0.01 with a non-augmented training dataset. The resulting model cannot correctly detect water sprouts, as shown in Figure 8 (a). The number of objects detected using this model is tiny; thus, the number of water sprouts detected at most was only 20%. The resulting model's performance on *T1* shows a low F1-score with a downward trend and an increase in the minimum confidence threshold, where the highest F1-score is 0.333.

In the second training scenario (*T2*), with an *LR* of 0.01 and an augmented dataset, the number of objects that the model could detect increased significantly (Figure 8 (b)). However, of all objects detected using this model, an average of 57% are *FP* detections; thus, the precision value tends to be smaller than the recall value. By using this model, the highest F1 score obtained is 0.543. It also shows that the performance improves with the minimum confidence threshold increase.

In the third scenario (*T3*) with *LR* 0.001 and without data augmentation, the model can detect water sprouts well (Figure 8 (c)). However, the number of objects the model detects is less than the number of water sprouts in the test dataset, indicating many *FN* detections. A higher minimum confidence threshold used results in a higher number of *FN* detections, which causes the model's performance to deteriorate along with the minimum confidence threshold used. From Figure 9, it can be seen that the resulting highest F1-score is 0.692, better than the model trained using *LR* 0.01.

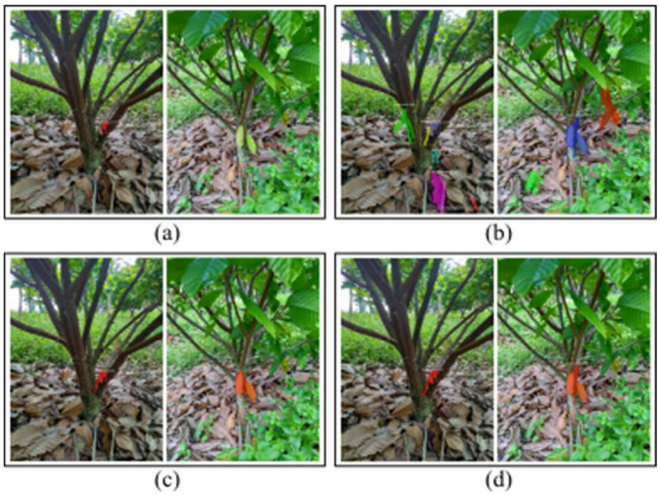

Fig. 8 Evaluation results example of (a) *T1*, (b) *T2*, (c) *T3*, (d) *T4*

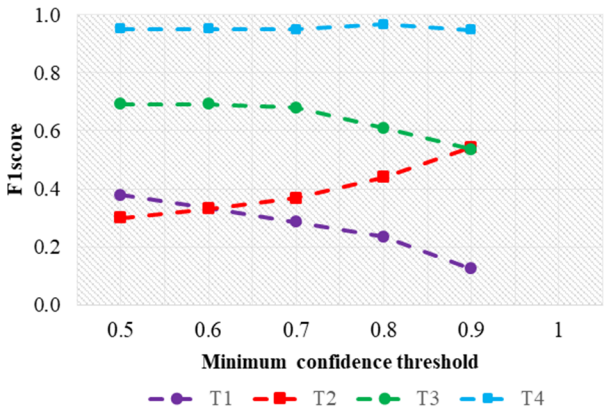

Fig. 9 Model performance based on F1score

In the last scenario (*T4*), with an *LR* of 0.001 and an augmented dataset, the highest F1-score performance is 0.966. The model generated in the fourth scenario is the best at detecting water sprout objects (Figure 8 (d)) compared to the three models used in this experiment. TP detection results are very high, above 90%, and the resulting detection error is relatively small, about 5%.

IV. Conclusion

This research performs the detection of water sprouts using the Mask R-CNN deep learning method. This paper performs data training using four scenarios by combining augmentation techniques and learning rate configuration. The Mask R-CNN model trained using data augmentation with a learning rate of 0.001 can detect water sprouts with the highest F1-score of 0.966. Data augmentation can improve the model's ability to

detect objects, and models without data augmentation suffer a lot of error detection. This research will be developed to create a water sprout-cutting robot in the future.

ACKNOWLEDGMENT

Universitas Hasanuddin funds this research through LPPM in PDU scheme 2020.

REFERENCES

[1] E. Emma, "Can water sprouts and suckers be prevented on trees?," Universitas of New Hampshire Extension, Feb, 26, 2021. [Online]. Available: https://extension.unh.edu/blog/2021/02/can-water-sprouts-suckers-be-prevented-trees.

[2] S. Abo-Hamed, H. A. Collin, and K. Hardwick, "Biochemical and Physiological Aspects of Leaf Development in Cocoa (theobroma Cacao L.)," *New Phytol.*, vol. 95, no. 1, pp. 9–17, 1983, doi: 10.1111/j.1469-8137.1983.tb03463.x.

[3] A. D. Mckelvie, "Cherelle Wilt of CacaoI. POD Development and ITS Relation to Wilt," *J. Exp. Bot.*, vol. 7, no. 2, pp. 252–263, Jan. 1956, doi: 10.1093/jxb/7.2.252.

[4] P. A. Sleigh, H. A. Collin, and K. Hardwick, "Distribution of assimilate during the flush cycle of growth in Theobroma cacao L.," *Plant Growth Regul.*, vol. 2, no. 4, pp. 381–391, Dec. 1984, doi: 10.1007/BF00027297.

[5] L. Diby, J. Kahia, C. Kouamé, and E. Aynekulu, "Tea, Coffee, and Cocoa," in *Encyclopedia of Applied Plant Sciences (Second Edition)*, B. Thomas, B. G. Murray, and D. J. Murphy, Eds. Oxford: Academic Press, 2017, pp. 420–425. doi: 10.1016/B978-0-12-394807-6.00179-9.

[6] G. Komitov, I. Mitkov, V. Harizanov, N. Neshev, and M. Yanev, "Justification of Agrotechnical Indicators of Agrorobot," in *2020 7th International Conference on Energy Efficiency and Agricultural Engineering (EE AE)*, Nov. 2020, pp. 1–5. doi: 10.1109/EEAE49144.2020.9279046.

[7] Y. Liu, X. Ma, L. Shu, G. P. Hancke, and A. M. Abu-Mahfouz, "From Industry 4.0 to Agriculture 4.0: Current Status, Enabling Technologies, and Research Challenges," *IEEE Trans. Ind. Inform.*, vol. 17, no. 6, pp. 4322–4334, Jun. 2021, doi: 10.1109/TII.2020.3003910.

[8] K. Jha, A. Doshi, P. Patel, and M. Shah, "A comprehensive review on automation in agriculture using artificial intelligence," *Artif. Intell. Agric.*, vol. 2, pp. 1–12, Jun. 2019, doi: 10.1016/j.aiia.2019.05.004.

[9] B. Narayanavaram, E. M. K. Reddy, and M. R. Rashmi, "Arduino based Automation of Agriculture A Step towards Modernization of Agriculture," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Nov. 2020, pp. 1184–1189. doi: 10.1109/ICECA49313.2020.9297546.

[10] V. Puranik, Sharmila, A. Ranjan, and A. Kumari, "Automation in Agriculture and IoT," in *2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, Apr. 2019, pp. 1–6. doi: 10.1109/IoT-SIU.2019.8777619.

[11] T. W. Cenggoro, A. Budiarto, R. Rahutomo, and B. Pardamean, "Information System Design for Deep Learning Based Plant Counting Automation," in *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*, Sep. 2018, pp. 329–332. doi: 10.1109/INAPR.2018.8627019.

[12] N. O. S. Matthew *et al*., "Robotic Automation in Agriculture," *International Journal of Trend in Research and Development*., vol. 8, no. 3, pp. 381-384, Jun 2021.

[13] T. Veramakali *et al*., "Smart Agricultural Management using IoT Based Automation Sensors," *International Journal of Recent Technologyand Engineering (IJRTE),* vol.8, no. 6, March 2020.

[14] L. Zhang, C. Xia, D. Xiao, P. Weckler, Y. Lan, and J. Lee, "A leaf vein detection scheme for locating individual plant leaves," in *2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT)*, Sep. 2018, pp. 1–4. doi: 10.1109/ICT-ROBOT.2018.8549901.

[15] N. M. Yusoff, I. S. Abdul Halim, N. E. Abdullah, and A. A. Ab. Rahim, "Real-time Hevea Leaves Diseases Identification using Sobel Edge Algorithm on FPGA: A Preliminary Study," in *2018 9th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, Aug. 2018, pp. 168–171. doi: 10.1109/ICSGRC.2018.8657603.

[16] J. Qi, D. Xie, L. Li, W. Zhang, X. Mu, and G. Yan, "Estimating Leaf Angle Distribution From Smartphone Photographs," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1190–1194, Aug. 2019, doi: 10.1109/LGRS.2019.2895321.

[17] Y. Chen, S. Baireddy, E. Cai, C. Yang, and E. J. Delp, "Leaf Segmentation by Functional Modeling," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2019, pp. 2685–2694. doi: 10.1109/CVPRW.2019.00326.

[18] Z. Wang, K. Wang, F. Yang, S. Pan, and Y. Han, "Image segmentation of overlapping leaves based on Chan–Vese model and Sobel operator," *Inf. Process. Agric.*, vol. 5, no. 1, pp. 1–10, Mar. 2018.

[19] M. A. Kutlugün, Y. Sirin, and M. Karakaya, "The Effects of Augmented Training Dataset on Performance of Convolutional Neural Networks in Face Recognition System," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2019, pp. 929–932. doi: 10.15439/2019F181.

[20] A. S. Paste and S. Chickerur, "Analysis of Instance Segmentation using Mask-RCNN," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, Jul. 2019, vol. 1, pp. 191–196. doi: 10.1109/ICICICT46008.2019.8993224.

[21] Y. Song and Z. Lin, "Species recognition technology based on migration learning and data augmentation," in *2018 5th International Conference on Systems and Informatics (ICSAI)*, Nanjing, Nov. 2018, pp. 1016–1021. doi: 10.1109/ICSAI.2018.8599361.

[22] S. Park, S. Lee, and J. Park, "Data augmentation method for improving the accuracy of human pose estimation with cropped images," *Pattern Recognit. Lett.*, vol. 136, pp. 244–250, Aug. 2020, doi: 10.1016/j.patrec.2020.06.015.

[23] A. Almutairi and M. Almashan, "Instance Segmentation of Newspaper Elements Using Mask R-CNN," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec. 2019, pp. 1371–1375. doi: 10.1109/ICMLA.2019.00223.

[24] M. Z. Islam et al, "Static Hand Gesture Recognition using Convolutional Neural Network with Data Augmentation," in 8th International Conference on Informatics, Electronics and Vision (ICIEV), Washington, USA, April 2019, doi:10.1109/ICIEV.2019.8858563.

[25] N. Maulidyah, Indrabayu, and I. S. Areni, "Water Sprouts Detection of Cacao Tree Using Mask Region-based Convolutional Neural Network," in *2020 27th International Conference on Telecommunications (ICT)*, Oct. 2020, pp. 1–5. doi: 10.1109/ICT49546.2020.9239443.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.322.

[27] S. Li, M. Yan, and J. Xu, "Garbage object recognition and classification based on Mask Scoring RCNN," in *2020 International Conference on Culture-oriented Science Technology (ICCST)*, Oct. 2020, pp. 54–58. doi: 10.1109/ICCST50977.2020.00016.

[28] J. Shi, Y. Zhou, and W. X. Q. Zhang, "Target Detection Based on Improved Mask Rcnn in Service Robot," in *2019 Chinese Control Conference (CCC)*, Jul. 2019, pp. 8519–8524. doi: 10.23919/ChiCC.2019.8866278.

[29] M. Bizjak, P. Peer, and Ž. Emeršič, "Mask R-CNN for Ear Detection," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2019, pp. 1624–1628. doi: 10.23919/MIPRO.2019.8756760.

[30] M. A. Malbog, "MASK R-CNN for Pedestrian Crosswalk Detection and Instance Segmentation," in *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Dec. 2019, pp. 1–5. doi: 10.1109/ICETAS48360.2019.9117217.

[31] X. Zhang, G. An, and Y. Liu, "Mask R-CNN with Feature Pyramid Attention for Instance Segmentation," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, Beijing, China, Aug. 2018, pp. 1194–1197. doi: 10.1109/ICSP.2018.8652371.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *ArXiv13112524 Cs*, Oct. 2014, [Online]. Available: http://arxiv.org/abs/1311.2524.

[33] K. S. Htet and M. M. Sein, "Toddy Palm Trees Classification and Counting Using Drone Video: Retuning Hyperparameter Mask-RCNN," in *2021 7th International Conference on Control, Automation and Robotics (ICCAR)*, Apr. 2021, pp. 196–200. doi: 10.1109/ICCAR52225.2021.9463466.

[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *ArXiv151203385 Cs*, Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385.

[35] T. Liu, M. Chen, M. Zhou, S. S. Du, E. Zhou, and T. Zhao, “Towards Understanding the Importance of Shortcut Connections in Residual Networks,” *ArXiv190904653 Cs Math Stat*, Nov. 2019, [Online]. Available: http://arxiv.org/abs/1909.04653.

[36] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” *ArXiv161203144 Cs*, Apr. 2017, [Online]. Available: http://arxiv.org/abs/1612.03144.

[37] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” *ArXiv14050312 Cs*, Feb. 2015, [Online]. Available: http://arxiv.org/abs/1405.0312.