Hybrid "Faster R-CNN" and "YOLO-v5" for Ripe Jackfruit Detection, from a Dataset Collected in Danang City, Vietnam

Ha Huy Cuong Nguyen [a,*] and Trung Hai Trinh[b]**

[a]Software Development Centre - The University of Danang, Danang City, Vietnam
Email: nnhcuong@sdc.udn.vn
[b]Vietnam - Korea University of Information and Communication Technology, The University of Danang, Danang City, Vietnam
Email: tthai@vku.udn.vn

Abstract: *The YOLO and Faster R-CNN models have high accuracy rates for identifying Ripe Jackfruits, but there are still some issues to resolve. Some Ripe Jack fruits were recognizable in YOLO, but not in Faster R-CNN. In this paper, we propose a method that combines Faster R-CNN and YOLO-v5 improved to detect ripening Jackfruit. Test case studies were conducted using 50,000 training and test images, respectively, with a hybrid model developed between YOLO-v5 improved and Faster R-CNN. The combined model achieved improved object detection accuracy and reduced training time, particularly for spoiled Jackfruit, which had a higher accuracy rate than previously tested individual models. Our proposed model achieved an accuracy of almost 99.8% in image classification for detecting and identifying different stages of Jackfruit, including semiRipe, unripe, ripe, and spoiled Jackfruit. This achievement resulted in a significant improvement in both the loss values and accuracy rates, making it a promising approach for identifying ripe Jackfruit with a high precision. Additionally, it reduced the training time compared to models proposed by previous researchers. The proposed model also improved performance based on the loss function and accuracy, and we also analyzed the number of training datasets. A test set of 50,000 images for each class, shows that the classifier has an accuracy of 99.5%*

Keywords: Feature Extraction, Convolutional Neural Network (CNN), Anchor Boxes, Non-Maximum Suppression (NMS), Intersection over Union (IoU), Transfer Learning, Loss Function

1. Introduction

Agriculture automation is rapidly expanding worldwide, driven by the Fourth Industrial Revolution's technological advancements. In Vietnam, a growing focus on innovation and investment in farming practices accompanies this trend, including digital transformation, global integration, and self-seeking investment opportunities. Ensuring increased productivity, economic efficiency, output and the quality of clean products, while maintaining nutrition is a priority for the whole society, with state-funded research teams developing unique solutions through multiple trials, to meet market demands. However, Vietnam's agricultural model still lacks sufficient technical solutions to enhance artificial intelligence's application in production.

To address this gap, recent advances in object detection have shown promising results in driving solutions. For instance, convolutional neural networks and rational proposal methods have been successful in driving recent advances in object detection, with Fast R-CNN achieving near real-time qualities by bypassing the time spent on region suggestions. Similarly, improved versions of Yolo-v4 and v5 extract high-level features from lower layers and use a grid of cells containing anchor boxes to predict confidence scores, conditional probabilities, and bounding boxes for each anchor box.

Faster R-CNN, in contrast, uses a Region Proposal Network (RPN) to find region proposals instead of the selective search algorithm, and the region-of-interest (RoI) composite subcaste uses maximal aggregation to convert features within an RPN-recommended region into a small dot plot.

Another model that has shown promising results in detecting objects is transferred deep learning, which reduces training time and image figures. In this paper, we propose a new method for enhancing Vietnam's agricultural model using these and other approaches. Specifically, we present our proposed method in Section 3, followed by our research results and discussion in Section 4. We conclude the paper by summarizing our findings and outlining future research directions in Section 5.

2. Literature Review

The R-CNN model, designed for detecting Ripe Jackfruit objects, was pioneered by two research groups led by Nguyen et al [1]. To precisely detect and identify objects from proposed regions, R-CNN integrates a selective search strategy, forming the foundation for the Fast R-CNN model [2]. However, this neural network-based object detection technique demands significant processing time. To streamline object detection with the Fast R-CNN model, previous studies have suggested a technique: conducting bounding box labeling of the image (also known as the region of interest) and iteratively refining until the desired region is detected, enabling bounding boxes to be applied [3][4].

In this article, we propose leveraging the selective search system in an enhanced Faster R-CNN model to process reference ranges by grouping bounding box areas into sets of three objects or more. Presently, several studies propose this method based on the R-CNN model, employing CNN and SVM to classify objects in input images [1][3]. However, previous studies did not achieve high object detection accuracy, and the detection time was slower compared to the Faster R-CNN method [1][16].

Researchers have developed a novel computer vision technique termed Mask R-CNN (Mask Region Convolutional Neural Network) as a modification of widely used Faster R-CNN and FPN models. Unlike Faster R-CNN, Mask R-CNN employs the ROIAlign layer instead of the ROIPooling layer and integrates a mask head. This technique utilizes simplified network structures throughout the network and employs a fully convolutional network for the mask head to enhance efficiency. These advancements could have significant implications across various applications, from medical imaging to self-driving vehicles [24]. Mask R-CNN can detect Jackfruit in images if it has been trained on a dataset inclusive of jackfruit images. This state-of-the-art object detection model can locate and identify objects of various classes within an image. However, for the model to specifically detect Jackfruit, it must have been trained on a dataset including Jackfruit as a class. Compared to the Faster R-CNN model, employing the Mask R-CNN model may not be ideal for certain applications due to the challenging process of regularly training the dataset before accurate recognition of objects, such as Jackfruit in non-growing environments or images captured by cameras, can be achieved [7][8][18][20].

In their study, Kuznetsova et al. found that the YOLO-v5 model accurately detected the ripeness stage of apples without requiring additional techniques [4]. They evaluated the models' performance using algorithms on the PASCAL Visual Object Classes (VOCs) dataset, highlighting the limitations of previous models due to limited data source libraries. This drove their research, particularly in enhancing jackfruit data in Vietnam [1][10]. Arun and Umamaheswara proposed an automatic method for recognizing and counting fruits of varying colors and complex shapes based on multiple images [16]. Sharma proposed utilizing deep learning models to detect and identify the ripeness stage of Jackfruit for timely harvesting and care, achieving an accuracy of over 95% [24]. Zhang suggested subtracting overlapping edges from the box object score and considering the boundary of the box, achieving a recall of 0.80 and a precision of 0.88 on test images [15]. Gorshick proposed a fast convolutional neural network-based model for object recognition (Fast R-CNN), which was approximately nine times faster than standard R-CNNs for training deep VGG-16 networks [12]. Additionally, Nguyen et al. used YOLO-v5 for better recognition of maritime objects, leveraging a public dataset with annotated videos and providing a unique training opportunity for deep neural networks (DNNs) in maritime object recognition [14].

Wang et al. utilized computer vision to rapidly, automatically, and accurately detect and register apples from sequential input images [13]. Artificial lighting during the night can limit natural light and pose challenges for fruit recognition systems, including those based on deep learning, as highlighted in previous studies [2][14]. In some studies, transfer learning was applied to reduce the number of parameters and computation costs for fruit recognition, where the target dataset was comparable to the training dataset in size and scope [22][25]. Upon reviewing the research by Nguyen and colleagues, differences were identified in the algorithms used, computation time, and solution performance [1][14][17].

3. Methodology

The approach outlined in this paper stems from our observation that YOLO-v5 can detect three jackfruit types in images ripe, unripe, semiRipe and spoiled Jackfruit while Faster R-CNN fails to do so [14][24]. Conversely, YOLO-v5 might miss certain Jackfruit detectable by Faster R-CNN [25]. To address these shortcomings, our proposed method integrates Faster R-CNN and YOLO-v5 into a unified network capable of detecting all three jackfruit types. If one network fails to detect a jackfruit type, the other network steps in for detection. The working diagram of our proposed method illustrates the various activities involved, with rounded rectangles denoting each activity and rectangles delineating input or output data.

In Figure 1, the rounded rectangles depict activities. For instance, a rectangle labeled "Train Faster R-CNN and YOLO-V5" signifies the training process of an input image dataset. A rounded rectangle labeled "Take an image" represents the testing activity involving images captured by a camera. The rounded rectangle with the text "Detect Jackfruit in the image by YOLO-V5 improved" illustrates the action of detecting objects in an image using the improved YOLO-V5.

Rectangles describe the input or output data of an operation. For example, a rectangle labeled "Faster R-CNN's bounding boxes and confidence scores" describes the contents of bounding boxes and confidence scores generated by Faster R-CNN.

Furthermore, the rounded, green rectangles represent the object detection operations of the Faster R-CNN model, while the yellow rectangle represents the object detection operations of the YOLO-V5 model. A small pink circle denotes the start of an event, while a small circle with an outer circle represents the end. The red regular trapezoid indicates the comparison of conditions for implementation.

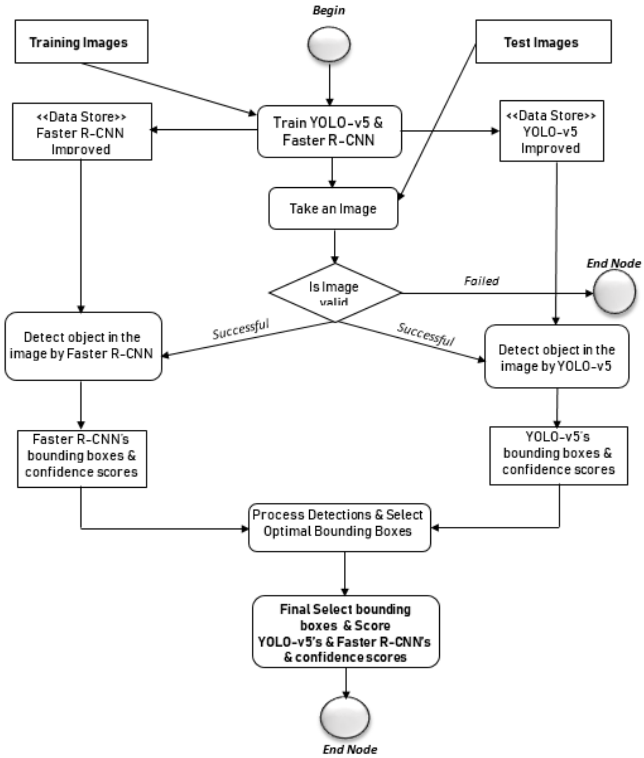


Figure 1. *The proposed method combines the Faster R-CNN and YOLO-v5 improved algorithms*

Utilizing the improved YOLO-V5 toolkit, this article presents technical solutions for the user-specific classification of Jackfruit. The jackfruit is detected and classified using a method proposed by us. The detection of Jackfruit is followed by classification into unripe, ripe, semiRipe and spoiled categories.

Algorithm 1: An algorithm that combines the Faster R-CNN model with the improved YOLO-v5 model

Input: A dataset containing images of semiRipe Jackfruits, unRipe Jackfruits Ripe Jackfruits and spoiled Jackfruit

Output: A well-trained network

Step 1. Preprocess the input image data.

Step 2. Load the training and validation sets into the algorithm.

Step 3. Perform data augmentation on the training set to increase its size and diversity.

Step 4. Apply the Faster R-CNN model to detect small objects in the image.

Step 5. Utilize the improved YOLO-v5 model to detect objects in the image and enhance processing efficiency.

Step 6. Combine the outputs of the Faster R-CNN and improved YOLO-v5 models to obtain the final object detection results.

Step 7. For each epoch (m) from 1 to T:

- Extract features through the Conv(CSPDarket) neural network.
- Generate the attention weight map using the Faster R-CNN and improved YOLO-v5.
- Concatenate the input feature map and output attention weight of the Faster R-CNN and improved YOLO-v5.
- Train the model using the stochastic gradient descent (SGD) algorithm.
- Evaluate the performance of the model M(m) using validation with the improved loss function.

Step 8. Save the optimal model with the highest mean average precision (mAP) in T epochs.

Step 9. Load the testing set into the algorithm.

Step 10. Load the optimal model and evaluate the object detection performance on the testing set.

Step 11. Post-process the object detection results to eliminate any duplicate or overlapping detections.

Jackfruit, including unRipe, semi-Ripe, ripe and overRipe variants, were identified using the YOLO-v5 algorithm as described in [15]. Our algorithm effectively detects and classifies Jackfruit through deep convolutional neural networks. Various data augmentation techniques were employed for different jackfruit types in our study. The utilization of neural networks eliminates the need for preprocessing. Our framework incorporates two connected layers and three convolutional layers. By training on extensive datasets of images, the network learns optimal features without requiring preprocessing. We extensively tested the proposed network across a range of real-life scenarios, and experimental results demonstrate its high accuracy. The combination of image enhancement techniques and filters has resulted in improved classification and validation accuracy. However, the current challenge lies in establishing an automated system for sorting, counting, detecting rotten Jackfruit, and grading multiple Jackfruit, as previous studies only featured datasets with a single fruit type. Future research can address this issue by using the Faster R-CNN model to enhance jackfruit quality prediction and grading in smart agriculture [7][8].

A deep learning system can predict the rate of jackfruit spoilage, enabling timely sales and benefits for both the seller and the customer. This system can also contribute to the country's economy.

3.1. Faster R-CNN

In the paper, we imported the input jackfruit image used in the training process into these integration classes. In this work, the problem of building an RPN score graph generates regional preferences for the resulting image. input Jackfruit and give their fair score. As for the Fast R-CNN model, starting with more advanced convolutional layers, it has thoroughly synthesized and processed the input images and created feature vector [17][18][23]. Our proposed model consists of two main components:

- The first component is the density map estimation model inherited from the VGG-16 model to extract the image attributes of the jackfruit growing area.
- The second is the weight map that has weighting at the pixel level for the honey map, the pixel with the Ripe Jackfruit image will have a greater weight and vice versa, the pixel with the most miniature Ripe Jackfruit or only semi-Jackfruit will have a small weight.

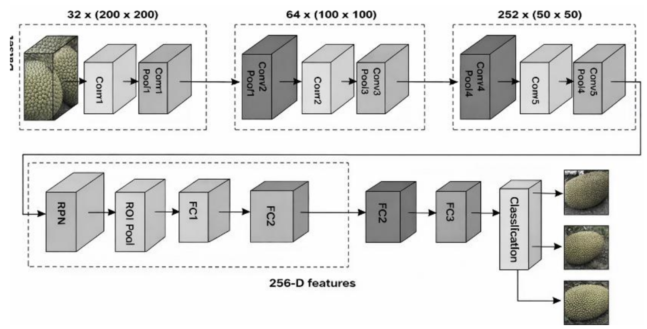


Figure 2. Proposed deep learning framework architecture of Jackfruit classification for Faster R-CNN

Multiplying the element between the density map and the weight map calculated the final density map. When combined like this, the weight map will weigh each pixel of the density map.

$$M(\{p_i\},\{T_i\}) = \frac{1}{Z_{cls}}\sum_i M_{cls}(p_i,p_i^*) + \lambda\frac{1}{Z_{reg}}\sum_i p_i^* M_{reg}(T_i,T_i^*) \qquad (1)$$

The calculation showed in Eq. (1), by:

- $M(\{p\},\{t_i\})$ *the* function loss of the Faster R-CNN model
- $M_{cls}(p_i, p_i^*)$ the loss index of classification
- $M_{reg}(t_i, t_i^*)$ regression loss
- T_i is a vector represented the parameterized coordinates of the predicted bounding box.
- T_i^* is that of the ground-truth box associated with a positive anchor

Proposed deep learning framework architecture of Jackfruit classification for Faster R-CNN

- Z_{cls}, and Z_{reg} and weighted by a balancing parameter λ
- i is the indicator of anchor in a minibatch.

p_i is the prognosticated probability of anchor

3.1.1. Ground truth map

Similar to previous approaches, we labeled the dataset of Ripe Jackfruit images ourselves. We generated a valid secret map by positioning the 2D Gaussian kernel at the marked points. The formula is as follows:

$$D^{XT} = \sum_{i=1}^{n} \mathrm{W}(z - z_i, \sigma) \qquad (2)$$

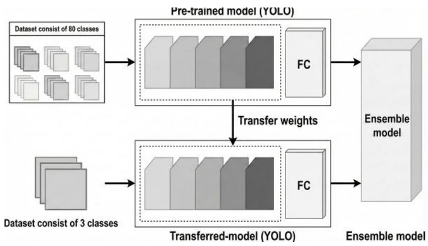


Figure 3. Augmentation and transformations for jackfruit images with Fast R-CNN

$$\delta = \beta \times \bar{d}_k \qquad (3)$$

Where:

- D^{XT} is the validation density map
- W Describes the use of Gaussian multiplication
- W_x^{XT} is the validation weight at point x with threshold α
- N is the number of images, C_i^{XT} is the true count of the i^{th} image
- $C_i(\alpha)$ is the count calculated using the weight map i^{th} with threshold α
- $\mathrm{W}_i^{XT} \log(P_i)$ is the validation weight map of the i^{th} image
- P_i is the probability of each pixel activated by the Sigmoid function

.

[RC1] megjegyzést írt:

- With standard deviation σ calculated by multiplying the distance from the point in question to k neighbors $\bar{d}_k$ and the parameter β

3.1.2. Validation weight map

As mentioned before, the confirmation weight chart will freight each pixel value of the secret chart. To be suitable for training the weight chart estimation model, we make a ground verify weight chart.

- Creating a confirmation weight chart is analogous
- Creating a secret chart
- The authentic secret chart

$$\forall x \in D^{XT}, \mathrm{W}_x^{XT} = \begin{cases} 1, \text{if } x > \alpha \\ 0, \text{if } x < \alpha \end{cases} \quad (4)$$

The density map pixel values are lower than the threshold α, indicating that pixels near the feature map will carry greater weight. Previous researchers utilized relatively small datasets of jackfruit-growing images for learning from scratch [19][20][23]. In larger-area images, the recognized image count is not high, simplifying their model. This is illustrated in Fig. 2 and Fig. 3. A final density map will be generated by combining the outputs of the two modules through element-wise multiplication. A loss function is employed to determine the value of α.

$$M_\alpha = \frac{1}{N}\sum_i^N \left| C_i(\alpha) - C_i^{XT} \right| \quad (5)$$

To train the proposed model, we use two loss functions, including:

1. Mean Square Error (MSE) for the density map estimation model
2. Binary Cross-Entropy loss function (BCE) for the weight map

$$M_2(\theta) = \frac{1}{2Z}\sum_{i=1}^{Z} \left\| D_i(\theta) - D_i^{XT} \right\|_2^2 \quad (6)$$

$$M_D = -\frac{1}{2Z}\sum_{i=1}^{Z} (D_i^{XT} \log(p_i) + (1 - D_i^{XT}) \log(1 - p_i)) \quad (7)$$

Therefore, during model training, the loss function will be the sum of the values of the two loss functions used for the density map and the weight map. This function is defined as follows:

$$M = M_2(\theta) + M_D \qquad (8)$$

3.1.3. Rating measures

To compare with previous methods, we use MAE and MSE:

$$MAE = \frac{1}{N}\sum_{1}^{N}\left|D_i - D_i^{GT}\right| \qquad (9)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(D_i - D_i^{GT})^2} \qquad (10)$$

Thus, the smaller the number of MAE and MSE, the higher the model accuracy. A part apprehensive RPN is proposed to replace the original RPN in the Faster R- CNN module, and a part apprehensive nonmaximum repression(NMS) is proposed to upgrade the final results as shown in Fig. 4.

3.2. YOLO-v5s Improved

The published research has provided an improved solution from the YOLO-v5 model. This study, based on a combination of two models, Faster R-CNN and YOLO-v5, aims to fully exploit the strengths of the two models to determine the ripening period of Jackfruit, with experiments being developed and implemented. In jackfruit regions in the central region of Vietnam. The initial experiment met the needs of the people when they wanted to determine the Ripe Jackfruit period, as well as the experiment at the pine-apple contractors [7][17][19][20].

The YOLO-v5 version was released. Within a few days, Glenn Jocher [21] announced the YO-LO-v5 model, and the author Glenn released a PyTorch-based version of YOLO-v5 with significant improvements.

The improved Yolo-v5 session used a combination of regression methods, bounding box handling, and computing probabilities of object recognition through the regression class. This improved result has increased the calculation speed, thus providing relatively high test results.

Regarding the technical features of the two proposed models from versions. Yolo-v5 and Faster R-CNN, the number of convolution layers is not the same. In these two models, our previous research also has improved solutions. As in the study [2] [6][14], we have improved the loss function.

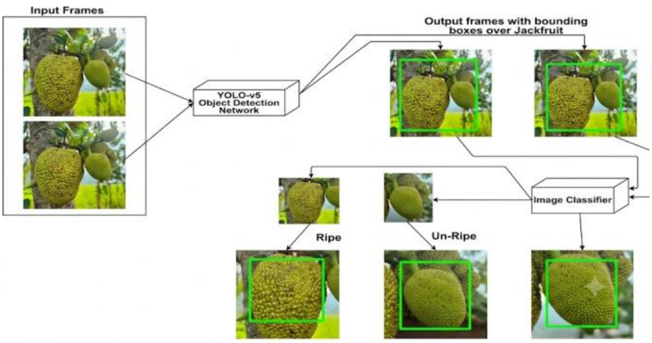


Figure 4. Recommend object detection with improved YOLO-v5

The jackfruit data image put into the test is divided into a matrix grid (SxS), with an improved version of Yolo-v5. However, with a parameter grid (SxS), when used for object recognition and detection, the process predicts whether an object with a center of gravity will fall into a certain grid. Then the technical method conducts the test by traversing the columns and rows of the grid matrix. Meanwhile, the test model is difficult and impossible to detect, or we performed the technical solution at a point of no reliability.

$$M_n = \frac{1}{2s_n} \sum_{m:r_{mn=1}} (x_m \mathrm{w}_n + b_n - y_{mn})^2 + \frac{\lambda}{2} |\mathrm{w}_n|_2^2 \ \lambda > 0 \qquad (11)$$

where s_n is the number of items that the n^{th} image:

$$s_n = \sum_{m=1}^{M} r_{mn} \qquad (12)$$

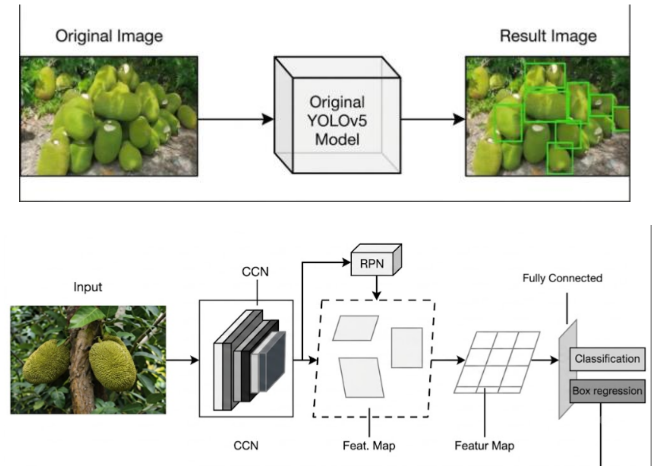


Figure 5. Test scenario with Ripe Jackfruit: Valid training and accuracy curves

It slides a convolutional subcaste network with completely connected layers over the feature map to induce region proposals. The small network is input by a spatial window from the feature map. There's a mapping between each sliding window and a lower-dimensional point. Following this, two completely connected layers have used a box- box-regression subcaste that produces the decoded coordinates of k anchor boxes and a box classification sub-layer that determines whether each offer is likely to be accepted or rejected. Figure 5 depicts the input dataset and the corresponding output after bounding the box.

Figure 6 depicts a confusion matrix, serving as a visual representation of the classification performance of a machine-learning model. The matrix presents the predicted classes juxtaposed with the actual classes in a tabular format. The accuracy of the model is computed by dividing the total number of correct predictions by all the predictions made. Specifically regarding jackfruit subclasses, the confusion matrix illustrates the classification outcomes for the Ripe Jackfruit, unRipe Jackfruit, semiRipe Jackfruit and spoiled Jackfruit categories [21].

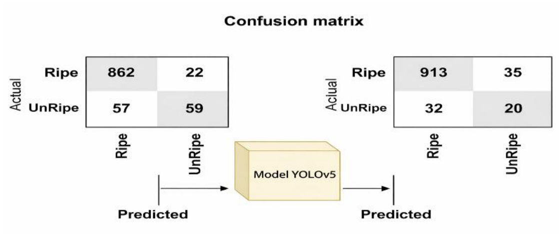


Figure 6. Shows the confusion matrix

A Fast R-CNN initiates by generating feature map from the input image utilizing convolutional and maximum-pooling layers. Experimental images are sourced from planting area images or cameras. The study implemented the image data enrichment process and the image enhancement step by transforming the features within an area recommended by the region proposal network (RPN) into a Feature vector. Subsequently, a region of interest (RoI) pooling sub-layer employs maximum pooling [20][22]. Following this, the feature map is converted into a feature vector using fully connected layers. The Improved YOLO-v5 [15][15] is employed in the study, and the formula is presented below:

$$M = M_1 + M_2 + M_3 \qquad (13)$$

where M_1, M_2, M_3 the symbols are the confidence loss function, the classification loss function, and the bounding box regression loss function, respectively.

The bounding box regression is:

$$M_3 = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{16}{\pi^4} \frac{(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^4}{1 - IoU + \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2} \qquad (14)$$

The calculation show in Eq. 14, where IoU is cross- road over. Between the verity bounding box and prognosticated bounding box.

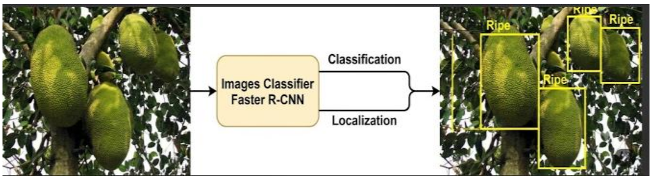


Figure 7. Ripe Jackfruit regions detected by Faster R-CNN

4. Results and Discussion

We utilized open-source ImageNet data along with images showcasing various ripening stages of Jackfruit. Additionally, we sourced jackfruit images from numerous agricultural farms in Danang City in Vietnam, yielding a comprehensive dataset of jackfruit images. The dataset comprised over 50,000 images collected from different locations and viewing angles, captured during the day and under diverse weather conditions. Our study conducted experiments within a collaborative environment [9][10][24][25].

4.1. Data Set Description

Jackfruit, a popular tropical fruit consumed worldwide, has garnered increased attention in recent years for its automatic detection and classification through computer vision and machine learning techniques. However, the availability of large and diverse jackfruit datasets remains limited. In this proposal, we introduce a new dataset comprising jackfruit images sourced from an orchard in Vietnam. This dataset serves as a valuable resource for tasks such as object detection and classification.

To collect the dataset, we will capture high-quality images of Jackfruit using a high-resolution camera. Images will be taken under various lighting conditions and from different angles to ensure dataset diversity. The dataset will encompass images depicting Jackfruit at different growth stages, including unripe, ripe, and spoiled stages. Annotation of the dataset will be accomplished through a combination of manual and automated methods. A subset of images will be manually annotated with bounding boxes and labels for object detection and classification, serving as training data for training an object detection model to automatically annotate the remaining images.

The objectives of this proposal are as follows:

- Collect a new dataset of jackfruit images sourced from an orchard in Vietnam.

- Annotate the dataset with bounding boxes and labels for object detection and classification tasks.
- Make the dataset publicly available for researchers and practitioners to utilize in their work.

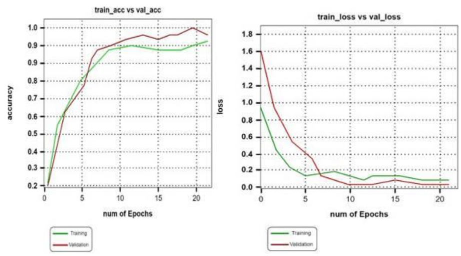
We anticipate gathering a new dataset comprising at least 50,000 jackfruit images from an orchard in Vietnam. This dataset will be annotated with bounding boxes and labels for object detection and classification tasks, and it will be made publicly accessible to the research community. This initiative will enable researchers and practitioners to develop and evaluate new algorithms for jackfruit detection and classification.

The proposed dataset of jackfruit images will serve as a valuable resource for the computer vision and machine learning community. The availability of a large and diverse dataset of jackfruit images will empower researchers and practitioners to develop more accurate and robust algorithms for detecting and classifying Jackfruit in images.

Based on the data collected from Fig. 6 and Fig. 7, reasons for misclassifications have been elucidated, legitimizing the errors, and the error rate has been computed over five training cycles at a learning rate of 0.0005. We achieved a precision rate of 99%. The proposed training process was based on the PyTorch deep learning library, utilizing a VGG-16 model pretrained with various fruits. The confusion matrix for the trained benchmark model is provided below. A disarray grid assesses the presentation of the prepared model on an informational index with realized objective classes.

4.2. Simulation Results

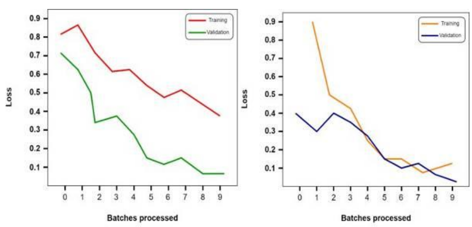
CNN divides recommendations based on regions while maintaining recommendations by region and aggregating ROIs, as with general R-CNN. Faster R-CNN maintains a common R-CNN frame: the CNN processes the entire input image first and then splits the recommendations. The Region Proposal Network speeds up the calculation of Region proposals. YOLO processes the divided image using CNN after the whole image is divided. Another advantage of YOLO is that it eliminates sliding windows and suggestions, as well as dividing input images into grids. For decision-making, each grid has a Confidence IoU mechanism with class probabilities. However, the faster R-CNN has nine anchor boxes instead of five in YOLO-v5, while YOLO-v5 re-uses anchor boxes. Faster R-CNN does not support dimensional clusters, powered by the improved YOLO-v5.



(a) Faster R-CNN (b) YOLO –V5 improved

Figure 8. Compare the improved YOLO-v5 accuracy function and Faster R-CNN on a training dataset of 10,000 images with num of Epochs

To validate the proposed model's efficacy and efficiency, we conducted tests with two individual models before proceeding with the second model. We restricted the evaluation parameters to include precision value and loss function. Through the test scenarios, it became evident that selecting the Faster R-CNN model for real-time object recognition could be challenging. Determining the optimal number of layers, layer types, and neurons per layer posed significant difficulty. This section examines various network architectures to identify the most effective learning approach. We employed twenty epochs for training the network. As the number of training epochs increased, the training loss curve consistently decreased. Strategies such as overfitting prevention, class skipping, augmented data utilization, generalized architectures, and hyperparameter tuning all yielded beneficial outcomes. The initial test, as depicted in Figure 12 and Figure 13, achieved a remarkable accuracy of 99% with 20 epochs, indicating elimination of overfitting successful.



(a) Faster R-CNN (b) YOLO –V5 improved

Figure 9. Compare the improved YOLO-v5 loss function and Faster R-CNN on a training dataset of 10,000 images with num of Epochs

Moreover, YOLO-v5 offers superior object representation compared to other libraries, making it an ideal tool for applications requiring fast and powerful object detection. Given its exceptional advantages, it deserves promotion and widespread adoption. In the realm of machine learning, the primary upcoming challenge lies in the scope of the dataset. Optimal results hinge on access to appropriate training data. An evaluation and decision-making process is necessary to differentiate Ripe Jackfruits from unripe fruit. The input image dataset was trained using the YOLO-v5 model, with all layers refined and optimized. In the test scenarios, the study employed a training rate of 0.0005 for initial classes and 0.00046 for final classes. An accuracy of 99.8% was achieved after 60 minutes of training. For the further developed model, computation results from the confusion matrix indicate probabilities of precise assessment and estimation at 99% and 100%, respectively.

4.3. Comparative Analysis of Other State-of-the-Art Methods

In this section, we analyze the results of the study conducted on 50,000 images of Jackfruit. The experimental results are analyzed based on the evaluation of the improved loss function charts for two models: Faster R-CNN and Improved YOLO-v5.

Table 2 provides the following insights: When YOLO-v5 is utilized as the feature extraction network, the detection speed averages the fastest, but the Average Precision (AP) value is the lowest. This discrepancy arises because the YOLO-v5 network structure is relatively simple and can only extract limited residual features. However, when YOLO-v5 is improved and Faster R-CNN is employed as the feature extraction network, precision improves significantly due to the complex network structure's ability to extract more intricate residual features, thereby resulting in a lower false detection rate. The highest AP value and longest average detection time are achieved when YOLO-v5 is improved as the feature extraction network. Comparing YOLO-v5 improved to the baseline YOLO-v5, there is a reduction in the AP value by 3.16%, but there is also a decrease in the average detection time by 65.28ms. Given the experimental conditions and research objectives, YOLO-v5 was ultimately chosen as the feature extraction network.

Comparing the y-axis values of the two model graphs in Fig.7 and Fig.8, when examining the loss function values concerning Bounding Box Regression, it is evident that the percentage reduction achieved with our improved YOLO-v5 model is ten times smaller than that of the Faster R-CNN model. In both the Faster R-CNN and YOLO-v5 models for image-based object detection, an object in the image may generate multiple bounding boxes. To address this, we employed an improved technique based on using a non-maximum suppression function, which helps eliminate bounding boxes for the same detected object. Consequently, the bounding box ratio of the YOLO-v5 model is smaller, while the improved technique based on the Faster R-CNN model exhibits a higher rate. Our algorithms operate in sequential steps, sorting all bounding boxes based on a confidence score.

In Fig.7 and Fig.8, we can observe the bounding-box regression values based on the loss function. The X-axis represents the epoch scale of the model, with the Faster R-CNN model spanning 15,000 epochs, while the YOLO-v5 model we utilized spans 20,000 epochs. The y-axis depicts the percentage of time spent training in the process of locating the detected object.

Figure 10. The interface is designed to run on multiple platforms, including the web and mobile apps

Figure 8 displays the web and mobile app we developed for the Yolo-v5 improved and Faster R-CNN complex. The system takes an input image, either captured or from a camera, and conducts a bounding box to detect objects. In the example image, the system can simultaneously detect three types of Jackfruit: semiRipeJackfruit, unRipe Jackfruits, Ripe Jackfruits and spoiled Jackfruit.

Table 1.

Comparison the optimal time detect object YOLO-v5 and Faster R- CNN Improved

Backbone	Precision	Recall	AP	F1-score	Test -time (ms)
YOLO-v5	93	94	90	0.9	280
YOLO-v5 Improved	94	98	91	0.91	298
Faster R-CNN	90	92	89,3	0.98	402

Figure 11. Image of Ripe Jackfruit identification using YOLO-v5 improved model

Figure 12. Image of Ripe Jackfruit identification using Faster R-CNN im- proved model

Figure 9 showcases images of jackfruit identification in growing areas and gathering areas for sorting and packing before transportation to factories and supermarkets. The user-friendly Web/App product has received high praise from jackfruit farmers and is currently in use in supermarkets. Additionally, it is available as a free Android mobile application for farmers in Vietnam.

Table 1 presents the training results as follows: The Average Precision (AP) values of YOLO-v5 and YOLO-v5 improved are lower than that of the Faster R-CNN model. The Faster R-CNN model, being a lightweight object detection model, exhibits faster object detection but lower accuracy due to incomplete residual feature extraction. YOLO-v5, being a single-stage detection model, lacks a Region Proposal Network (RPN) and relies on direct classification of prediction boxes through feature information extraction, resulting in under-utilization of the backbone network's feature extraction ability. Although YOLO-v5 and YOLO-v5 improved demonstrate faster detection speeds than the Faster R-CNN model, they exhibit lower detection accuracy. In contrast, the Faster R-CNN model ensures higher detection accuracy and guarantees real-time detection.

Figure 10 illustrates an experimental image of jackfruit identification using the Web/App, which combines two improved models: YOLO-v5 and Faster R-CNN. The experiment results have been promising, as the system accurately and completely identifies objects, with almost absolute accuracy, in both images and videos. This solution proves highly effective for the Fourth Industrial Revolution and actively supports digital transformation in smart agriculture and high-tech agricultural models in agriculturally developed countries.

Making a fair comparison between object detection models is challenging. Comparing Figure 10, Figure 11, and Figure 12, we lack an exact answer regarding the best model. For detection applications focusing on Ripe Jackfruits, when we utilize a selection of Faster R-CNN and YOLO-v5 models, some images are detected by one model while the other model fails to detect them.

After testing various models, YOLO-v5 improved and YOLO-v5 demonstrate outstanding performance, as depicted in Table 2 and Table 3. Their precision values exceed 99%, recalls hover around 96%, and mean average precisions surpass 98.7%. However, the Faster R-CNN model exhibits the lowest mAP, and its detection performance is suboptimal. Compared to YOLO-v5s, YOLO-v5m, YOLO-v5l, and Faster R-CNN, YOLO-v5 improved significantly outperforms in terms of detection performance.

Figure 13 illustrates the performance comparison results of the improved method based on the loss function class between Faster R-CNN and YOLO-v5 improved models. In table 4 the test dataset comprises 10,000 images of three types of Jackfruit: semi-Jackfruit, unRipe Jackfruits, and Ripe Jackfruits. Experimental results indicate that the average difference in the loss function ratio between the two models is rather small, at 0.01%. As the value of k increases, the ratio difference remains minimal, and the graph demonstrates that the models exhibit similar performance, making it challenging to distinguish their advantages.

Table 2.
Compare performance between network models

Model	AP	F1-score	Test-time(ms)
YOLO-v5	67.50	0.70	194.75
YOLO-v5 Improved	75.20	0.74	178.45
Faster R-CNN	84.34	0.92	250.78

Table 3.
Comparison the optimal time detect object YOLO-v5 and Faster R-CNN Improved

Model	Recall	Precision	Test-times	MAG	Memory
YOLO-v5s	96	97	98	9.2	24%
YOLO-v5m	95	98	96	9.5	24.5%
YOLO-v5l	95.5	97.6	95.6	12.2	68%
YOLO-v5 Improved	95.6	96.4	95.3	27.6	269%
Faster R-CNN	97	98.5	98.6	29.8	327%
Faster R-CNN Improved	97.23	86.7	89.4	26.82	296%

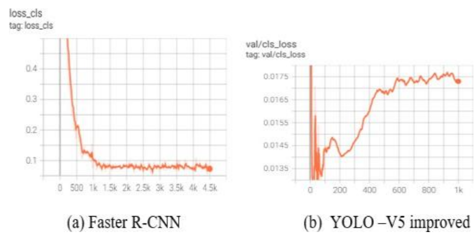


Figure 13. A comparison of the performance of YOLOv5 and Faster R-CNN

Therefore, in this paper, we have presented a hybrid method between the two models to address the requirements of the problem, aiming to strike a balance between accuracy and speed. In addition to selecting detection models, understanding the choices that affect model accuracy is crucial. Based on the chart analysis, we have observed the balance between the two algorithmic models. Although there exists a difference between them, it is deemed acceptable as it falls within the evaluation limits.

The utilization of a hybrid model combining both models encounters numerous difficulties and challenges in research and practical implementation. However, we have endeavored to implement and restrict some parameters in the comparison. Simulation results indicate that the overall loss was superior to that of the improved YOLO-v5 and Faster R-CNN, rendering better performance with minimal processing time. While both Faster R-CNN and YOLO employ CNNs as core components, they aim to enhance CNN-based recommendations. However, their frameworks differ significantly. Considering the broader context, the issue of preparing for loss is minimal, but the significance of achieving excellent results is high. The developed YOLO-v5 has been utilized to produce trained classifiers for images based on semi-Jackfruit and unRipe Jackfruits. These outcomes can assist farmers in identifying ripe or unripe produce, facilitating distribution to stores and global exportation. Farmers have greatly benefited from the results of this research. To further support the findings, a subset of the originally segmented data was utilized to validate the set of standard classes

5. Conclusions

In this work, we present effective solutions for identifying Jackfruit on rural homesteads in Vietnam, categorizing them as semi-Jackfruit and unRipe Jackfruits, and determining their quality (distinguishing between good and rotten Jackfruit). The motivation behind the article is to introduce a novel algorithm capable of identifying jackfruit at nine distinct stages.

To achieve this, we have devised an efficient approach that reduces the computational cost by employing the YOLO-v5 detection model to simultaneously detect and classify Jackfruit. Our study has successfully achieved the following objectives:

First, we propose a dataset comprising 50,000 jackfruit images for training and experimentation to detect Jackfruit at nine stages, extracted from on-site image data or cameras.

Second, the model is trained on approximately 20,000 images of semi-Ripe Jackfruits and 20,000 images of unRipe Jackfruits for each class.

Third, we have improved the training time significantly, reducing it to just a few hours for the entire set of 50,000 images using both Faster R-CNN and YOLO-v5 models.

And Last, we utilize a selection model employing the ensemble technique to enhance the overall accuracy of our proposed model, achieving an accuracy rate of approximately 94.2%.

Our proposed technique demonstrates flexibility as it can easily accommodate various classification tasks, such as disease detection in Jackfruit, effectively enhancing the model's capabilities.

Acknowledgment

This research is funded by Ministry of Education and Training under project number B2026-DNA-06

References

[1] Nguyen, H.H.C., Ho P.H., Nguyen, T.T Tran A.K., Vo T.H.,: An Enhanced Data Mining Classification based on Shuffled Frog Leaping Optimization, Acta Polytechnica Hungarica, Vol. 22, No. 5, 2025, 127 – 145.

[2] Khaing, Z. M., Naung, Y., and Htut, P. H., Development of con- trol system for fruit classification based on convolutional neural network, in Proceedings of the 2018 IEEE Conference of Rus- sian Young Researchers in Electrical and Electronic Engineer- ing (EIConRus), Moscow and St. Petersburg, Russia, 29 Jan.-1 Feb. (2018).

[3] Katalin Ferencz, József Domokos, Levente Kovács, Cloud Integration of Industrial IoT Systems. Architecture, Security Aspects and Sample Implementations; Acta Polytechnica Hungarica; Vol. 21, No. 4, 2024; 7 – 28.

[4] Nguyen, H.H.C., Luong, A.T., Trinh, T.H., Ho P.H., Meesad, P., Nguyen, T.T: Intelligent Fruit Recognition System Using Deep Learning. In: Meesad P., Sodsee D.S., Jitsakul W., Tangwan- nawit S. (eds) Recent Advances in Information and Communica- tion Technology 2021. IC2IT 2021. Lecture Notes in Networks and Systems, vol 251. Springer, Cham (2021).

[5] Jun Hwa, K., Namho, K., Woon, Park P. Y.,Sun, W. C.: Object Detection and Classification Based on YOLO-v5 with Improved Maritime Dataset. In Journal of Marine Sci- ence and Engineering, J. Mar. Sci. Eng. 2022, 10, 377. https://doi.org/10.3390/jmse10030377, (2022).

[6] Kuznetsova, A., Maleva, T., and Soloviev, V.: Detecting Ap- ples in Orchards Using YOLO-v3 and YOLO-v5 in General and Close-Up Images. International Symposium on Neural Net- works, ISNN 2020: Advances in Neural Networks – ISNN 2020 pp 233–243, 27 November (2020).

[7] Bargoti, S. and Underwood, J.: Deep fruit detection in orchards; IEEE International Conference on Robotics and Automation (ICRA), Singapore, 3626-3633. doi: 10.1109/ICRA.2017.7989417, (2017).

[8] Nguyen, H.H.C., Nguyen, D.H., Nguyen, V.L., Nguyen, T.T.: Smart solution to detect images in limited visibility conditions based convolutional neural networks. In: Advances in Com- putational Collective Intelligence. ICCCI 2020. Communica- tions in Computer and Information Science, vol. 1287, pp. 641–650. Springer, Cham (2020). https://doi.org/10.1007/978-

3-030-63119-2-52.

[9] Kim,S. , Ji, Y. and Lee, K.: An Effective Sign Language Learn- ing with

Object Detection Based ROI Segmentation, second IEEE International Conference on Robotic Computing (IRC), Laguna Hills, CA, 330-333 (2018).

[10] Common Objects in Context. [online]. Available: http://cocodataset.org/

[11] Open Images Dataset V5. [online]. Available: https://storage.googleapis.com/openimages/web/index.html.

[12] Jocher, G., Nishimura, K., Mineeva, T., Vilariño,R., - YOLOv5 (2020).https://github. com/ultralytics/yolov5.

[13] Girshick, R., Fast R-CNN. arXiv:1504.08083, 2015

[14] Wendong, G.„ Yakun, L., Jing, Z. and Gang, J.: An improved Tiny YOLOv3 for real-time object detection; Systems science and control engineering: An Open Acess Journal, vol. 9, no. 1, 314–321 (2021).

[15] Nguyen, H.H.C., Trinh, T.H., Meesad, P., Nguyen, T.T: Im- proved YOLO object detection algorithm to detect ripe pineapple phase: Journal of Intelligent and Fuzzy Systems, vol. 43, no. 1, pp. 1365-1381, (2022).

[16] Zhang, Yan et al. ‘An Improved YOLO Network for Unopened Cotton Boll Detection in the Field’. Journal of Intelligent and Fuzzy Systems, vol. 42, no. 3, pp. 2193-2206, (2022).

[17] Arun, R. and Umamaheswari, S. Effective and Efficient Multi- crop Pest Detection Based on Deep Learning Object Detection Models.Journal of Intelligent and Fuzzy Systems, vol. Pre-press, no. Pre-press, pp. 1-19, (2022).

[18] Li, Junpeng et al. Deep Neural Network-based Real Time Fish Detection Method in the Scene of Marine Fishing Supervision. Journal of Intelligent and Fuzzy Systems, vol. 41, no. 3, pp. 4527-4532, (2021).

[19] Alexey, A. B.: Apple detection during different growth stages in orchards using the improved YOLO-V3 model. https: github.com/AlexeyAB/Yolo (2018)

[20] Joseph, R., Santosh, D., Ross, G., Ali, F.: You Only Look Once: Unified, Real-Time Object Detection; arXiv:1506.02640 [cs.CV], May (2016).

[21] Kim, K.,Kim, P., Chung, Y. and Choi,D., Multi-Scale Detector for Accurate Vehicle Detection in Traffic Surveillance Data, in IEEE Access, vol. 7, pp. 78311-78319, 2019, doi: 10.1109/AC- CESS.2019.2922479.

[22] Sindagi, V.A., Yasarla, R., Patel, V.M. JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method. IEEE Trans Pattern Anal Mach Intell. 2022 May;44(5):2594-2609.

doi: 10.1109/TPAMI.2020.3035969. Epub 2022 Apr 1. PMID: 33147141.

[23] Glenn, J., Alex, S., Jirka, B., Ayush, C., Tao, X., Liu, C.Y., Ab- hiram, V.,Tkianai, L., AWS, Supervise. ly and YouTube integra- tions, Volume 10,

2021.

[24] Horea, M., Mihai, O., Fruit recognition from images using deep learning, arXiv:1712.00580

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

[26] Sharma A. K. , Nguyen H. H. C., Bui T. X. ,Bhardwa S. and Van Thang D., "An Approach to Ripening of Jackfruit Fruit with Model Yolo v5," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-5, doi: 10.1109/I2CT54291.2022.9824067.

[27] Chu J. , Guo Z. and Leng L. , "Object Detection Based on Multi-Layer Convolution Feature Fusion and Online Hard Ex- ample Mining," in IEEE Access, vol. 6, pp. 19959-19967, 2018, doi:10.1109/ACCESS.2018.2815149.