Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2019~2028
e-ISSN: 2548-6861

A Hybrid Framework Based on YOLOv8 and Vision Transformer for Multi-Class Detection and Classification of Coffee Fruit Maturity Levels

Ahmad Subki [1]*, M. Zulpahmi [2], Bahtiar Imran [3]****
*Rekayasa Perangkat Lunak, Fakultas Teknologi Informasi dan Komunikasi, Universitas Teknologi Mataram
**Rekayasa Sistem Komputer, Fakultas Teknologi Informasi dan Komunikasi, Universitas Teknologi Mataram
ahmad.subki1992@gmail.com [1], fahmijorge04@gmail.com [2], bahtiarimranlombok@gmail.com [3]

Article Info

Article history:

Received 2025-07-31
Revised 2025-08-30
Accepted 2025-09-10

Keyword:

YOLOv8,
Vision Transformer,
Object detection,
multi-class classification of coffee fruits.

ABSTRACT

Detection and classification of coffee cherries based on maturity levels present a significant challenge in agricultural product processing systems, primarily due to the high visual similarity among classes within a single bunch. This study aims to develop a multi-class detection and classification system for coffee cherries by integrating YOLOv8 and Vision Transformer (ViT) as a classification enhancer. The initial detection process is conducted using YOLOv8 to identify and automatically crop coffee cherry objects from bunch images. These cropped images are then re-classified using the Vision Transformer to improve prediction accuracy. The training process was carried out with a learning rate of 0.0001, a batch size of 16, and epoch variations of 50, 100, and 150. Evaluation results demonstrate that the integration of YOLOv8 and ViT significantly improves classification accuracy compared to using YOLOv8 alone. At 100 epochs, the YOLOv8+ViT model achieved an accuracy of 89.52%, a precision of 90.43%, and a recall of 89.52%, outperforming the standalone YOLOv8 model, which only reached an accuracy of 75.44%. These results indicate that the Vision Transformer effectively enhances classification performance, particularly for visually similar coffee cherry classes. The integration of these two methods offers a promising alternative solution for improving image-based multi-class classification in agriculture and other domains involving complex visual objects.

I. INTRODUCTION

Coffee is one of Indonesia's major commodities [1], including in Sajang Village, Sembalun, East Lombok, which is renowned for its high-quality Arabica coffee [2]. Grown in highland areas with volcanic soil, coffee from this region possesses a distinctive flavor profile [3]. Optimal harvesting relies heavily on selecting cherries at the right maturity level; however, manual methods remain subjective, inefficient, and prone to human error. Therefore, an AI-based automated technology is needed to detect and classify the maturity stages of coffee bunches with high accuracy and efficiency [4]. Artificial Intelligence (AI) has been increasingly utilized in agriculture to support automation and improve production efficiency [5]. In this context, detection models based on YOLOv8 and Vision Transformer present an innovative solution for accurately and consistently identifying the maturity levels of coffee bunches. The combination of these two models enables the system to rapidly recognize objects while performing deep analysis of the visual characteristics of coffee cherries. This approach can be implemented to assist farmers in Sajang Village, Sembalun, in enhancing the quality of their harvests.

Several studies have been conducted to detect and classify coffee cherry maturity using various image processing and machine learning techniques. For example, study [6] successfully improved maturity detection using a YOLOv7-based approach combined with optimization techniques, achieving a mean average precision (mAP) of 80.1%. However, this model was limited to single-object classification and did not incorporate Transformer-based

architectures or Explainable AI (XAI) components to enhance system transparency.

Previous studies have explored various approaches for classifying the maturity stages of coffee cherries. For instance, study [7] compared the performance of YOLOv3 and YOLOv4 in detecting coffee cherry maturity levels (green, ripe, overripe), with YOLOv4 achieving a mean average precision (mAP) of 81%. However, the model struggled to distinguish green cherries from leaves due to color similarity and was not yet effective in handling multi-class classification within a single bunch. Meanwhile, conventional machine learning approaches such as Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN) were employed in study [8], which focused on the physical and chemical changes of coffee cherries. While these methods demonstrated high accuracy in laboratory settings, they are less practical for field implementation due to the need for additional measurements and the absence of direct object detection capabilities.

In the domain of Transformer-based applications, study [9] was among the first to implement ViT, DeiT, and Swin Transformer for classifying the maturity levels of Arabica coffee beans, with Swin Transformer achieving the highest accuracy of up to 84.75%. Its main advantage lies in the adoption of Transformer architectures, which remain rarely explored in this field. However, the dataset used was limited to Arabica coffee beans sourced from USK-Coffee and has not yet been tested on other coffee varieties or combined with hybrid methods for further accuracy enhancement.

On the other hand, study [10] employed the k-Nearest Neighbors (KNN) algorithm in conjunction with computer vision and an Arduino Mega microcontroller to detect defects in green coffee beans. The integration of visual features produced more accurate results compared to manual methods. Nevertheless, the use of MATLAB and the KNN approach rendered the system less scalable for industrial applications.

Furthermore, study [11] compared several versions of YOLO, with a custom-YOLOv8n achieving a mAP of 0.995 and a precision of 0.987. Although highly accurate, the study focused solely on detecting bean defects rather than assessing overall fruit maturity levels, and it did not incorporate Transformer-based classification support.

A non-destructive method was also introduced in study [12], using chromaticity mapping from digital images to characterize coffee fruit maturity stages. While offering an automated and environmentally friendly approach, the method is highly susceptible to lighting variations and image quality, which can reduce classification accuracy. In Toraja, study [13] utilized a Convolutional Neural Network (CNN) to classify Arabica coffee fruit maturity levels, achieving an accuracy of 98.75% based on 5-fold cross-validation. However, this study only focused on classifying individual fruits and did not incorporate realistic object detection using approaches such as YOLO. Finally, study [14] utilized colorimetry and clustering techniques to classify the ripeness of Robusta coffee cherries with high accuracy. Although effective under controlled conditions, the method lacks adaptability to varying lighting environments and does not leverage deep learning approaches that could enhance model generalization.

Building upon the limitations of previous studies that focused solely on individual coffee fruit detection and color-based classification, this study proposes a hybrid framework based on YOLOv8 and Vision Transformer (ViT) to detect and classify the maturity levels of coffee fruit bunches (green, yellow, red, and black). The dataset was collected from plantations in Sajang Village, Sembalun, with key challenges including natural lighting variations, occlusion by leaves and branches, and varying maturity stages within a single bunch. YOLOv8 is employed for object detection, while ViT is utilized to reinforce the classification results by leveraging its capability to extract and understand complex visual features more effectively.

II. METHODS

The following is a comprehensive and detailed explanation based on the provided methodological diagram:

A. Literature Review

In the initial stage, a comprehensive literature review was conducted to identify the most effective methods for detecting and classifying the maturity levels of coffee cherries. This review included an analysis of various deep learning techniques, particularly the YOLOv8 model for object detection and the Vision Transformer (ViT) for image classification. YOLOv8 is widely recognized for its high performance in real-time object detection with strong accuracy [6], [15], hile ViT excels in capturing global feature representations from images [9]. Additionally, model optimization techniques such as the Adam optimizer and CrossEntropy loss function were reviewed to enhance classification performance. The literature study also covered relevant works on data augmentation methods aimed at improving model generalization under varying lighting conditions and image variability.

B. Dataset Collection

Following the methodological review, the next step involved collecting an image dataset of coffee fruit bunches directly from plantations in Sajang Village, Sembalun, East Lombok. The dataset comprises images captured under diverse lighting conditions, angles, and variations in fruit maturity within a single bunch (green, yellow, red, and black). Image acquisition was performed using a high-resolution DSLR camera—Canon EOS 90D—to ensure high-quality visuals, thereby facilitating the model's ability to accurately recognize the visual features of the coffee bunches.

C. Pre-processing

The data preprocessing stage aims to enhance the quality of the dataset prior to training the model. Several steps were performed, including image resizing, pixel normalization, data augmentation, and format conversion to ensure compatibility with the employed deep learning models. Image resizing was carried out to standardize image dimensions, such as resizing to 224×224 pixels for ViT and 640×640

pixels for YOLOv8. Pixel normalization was applied by converting pixel values from the range [0, 255] to [0, 1] using the following formula:

$$X_{norm} = \frac{x - x_{min}}{x_{max} - x_{max}}$$

where X is the original pixel value, X_{min} is the minimum value (0), and X_{max} is the maximum value (255). In addition, data augmentation techniques such as rotation, flipping, and contrast adjustment were applied to improve the model's robustness against various lighting conditions and image capture angles.

D. Dataset Splitting

After preprocessing, the next step was to divide the dataset into three main subsets: training set, validation set, and testing set. This partitioning was performed using a 70:20:10 ratio, with 70% of the data used for training the model, 20% for validation, and 10% for testing. Stratified sampling techniques were employed to ensure that the class distribution remained balanced across all subsets. The dataset used in this study consisted of full images of coffee fruit bunches, which were utilized throughout the training, validation, and testing processes. In total, the dataset was divided into 600 images for the training set, 174 images for the validation set, and 85 images for the testing set. The images represented a wide range of lighting conditions, backgrounds, and fruit maturity levels within each bunch, covering five maturity categories: unripe, semi-ripe, ripe, overripe, and dry. This data partitioning can be formulated as follows:

$$D_{train,}D_{val,}D_{test} = split(D, P_{train,}P_{val,}P_{test})$$

with $P_{train} = 0.7, P_{val} = 0.2$ and $P_{test} = 0.1$. This technique is essential to prevent the model from being biased due to an imbalanced data distribution.

E. Model Development (YOLOv8 + Vision Transformer)

In this stage, a detection and classification model was developed using two main approaches: YOLOv8 for object detection and the Vision Transformer (ViT) for classifying the maturity levels of coffee cherries. YOLOv8 was employed to detect the locations of coffee cherries within an image, while ViT was utilized to classify their maturity levels based on features extracted from the cropped images. The YOLOv8 architecture comprises three main components—backbone, neck, and head—that enable efficient object detection. The development and experimentation of the proposed model were carried out using Python within the Google Colab Pro environment[11]. The Vision Transformer (ViT) was implemented as a post-prediction classification module, which classifies the maturity level of each coffee cherry based on the cropped regions obtained from YOLOv8. ViT operates by dividing each cropped image into small patches of 16×16 pixels, with each patch transformed into an embedding vector. These patch embeddings are then passed through a series of transformer encoders to generate a comprehensive image representation, which is subsequently used for classification of the fruit's maturity level. The embedding process in ViT is formulated as follows:

$$Z_0 = [X_1E; X_2E; \dots; X_N] + E_{pos}$$

where X_i represents the image patch, E is the embedding matrix, and E_{pos} denotes the positional encoding.

Self-Attention Layer in Transformer:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

F. Model Evaluation

After the training process, model evaluation is conducted to assess performance. Several metrics are employed to evaluate the detection capability of YOLOv8, including mean Average Precision (mAP), Precision, Recall, and F1-score. For the Vision Transformer (ViT) classification model, evaluation metrics include Accuracy, Confusion Matrix, and the Receiver Operating Characteristic - Area Under Curve (ROC-AUC). The formulas for Precision and Recall [11] are defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Meanwhile, the mean Average Precision (mAP) is computed as follows:

$$mAP = \frac{1}{N}\sum_{I=1}^{N} AP_i$$

where AP_i denotes the Average Precision for the i class. The evaluation is conducted on the test dataset to ensure that the model performs well on previously unseen data.

G. Analysis and Recommendations

The final stage involves analyzing the model evaluation results to provide recommendations for performance improvement. If the model exhibits overfitting, techniques such as dropout, additional data augmentation, or parameter fine-tuning can be applied. In cases where performance remains suboptimal, further exploration of model architectures or advanced training techniques may be considered. The results of this study can also be compared with alternative methods to identify the strengths and limitations of the proposed approach.

III. RESULTS AND DISCUSSION

A. YOLOv8 Training Results

The training process was conducted to develop an accurate and robust maturity detector for coffee cherries using the YOLOv8 architecture. During the training phase, the model

was trained for 100 epochs with specific optimization parameters, including a learning rate of 0.0001 (1e-4), a batch size of 16, and a data augmentation strategy. The training results are illustrated in Figure 1.

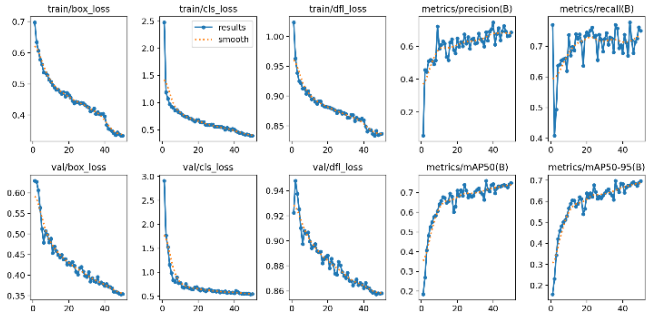

Figure 1. YOLOv8 Training Results

Figure 1 illustrates the training performance of the YOLOv8 model over 100 epochs with a learning rate of 0.0001 and a batch size of 16. The graphs indicate a consistent downward trend in loss metrics for both training and validation datasets. The train box loss, classification (cls) loss, and distribution focal loss (dfl) significantly decreased as the number of epochs increased, suggesting that the model effectively learned the data patterns. Lower loss values reflect a reduced discrepancy between the model's predictions and the ground truth across training iterations.

Furthermore, the precision and recall curves show a stable upward trend throughout the training process, reaching relatively high values above 0.7 by the end of the training. This indicates that the model is not only capable of detecting objects accurately (precision) but also able to identify most of the relevant objects in the images (recall). Meanwhile, the mAP50 and mAP50-95 metrics also exhibit substantial improvements over time. The mAP50 score exceeded 0.85 at epoch 50 and remained stable through epoch 100, while the more stringent mAP50-95 also demonstrated consistent growth, reflecting the model's reliable object detection performance across various Intersection over Union (IoU) thresholds.

Overall, these training results indicate that the YOLOv8 configuration used in this study successfully produced an optimal object detection model. The consistent decrease in loss and improvement in evaluation metrics on both training and validation sets suggest that the model did not suffer from overfitting or underfitting, and it generalized well on the unseen test data.

Figure 2 presents the Precision-Confidence Curve of the YOLOv8 object detection model applied to the coffee fruit dataset consisting of five ripeness classes: dry, overripe, ripe, semi_ripe, and unripe. This curve illustrates the relationship between the confidence threshold (X-axis) and precision (Y-axis) for each class. In general, the precision tends to increase with higher confidence thresholds, indicating that predictions with greater confidence levels are more likely to be accurate.

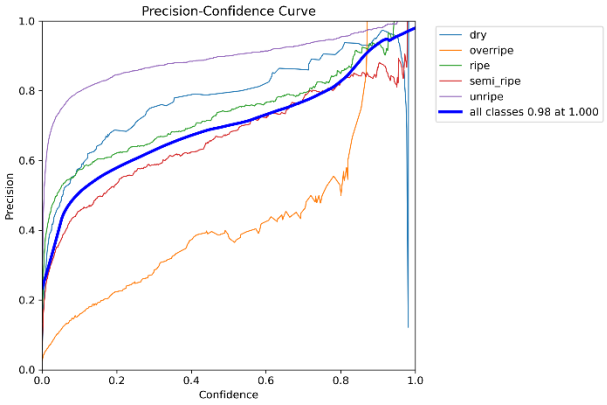

Figure 2. Precision-Confidence Curve (P-R Curve) of YOLOv8 Detection Results

Based on the results, the unripe class demonstrates the best performance, exhibiting high and stable precision values across almost the entire confidence range, followed by the dry, ripe, and semi_ripe classes. In contrast, the overripe class shows comparatively lower precision, particularly at confidence levels below 0.7, before experiencing a sharp increase approaching a confidence of 1.0. This suggests that the model has greater difficulty distinguishing overripe objects at low to moderate confidence levels, possibly due to visual similarity with other classes or a limited number of training samples.

The thick blue line in the graph represents the average precision across all classes, peaking at 0.98 at a confidence level of 1.0. This high value indicates that the model's predictions are highly reliable when it assigns maximum confidence. This curve serves as a crucial indicator for determining the optimal confidence threshold for deployment, by balancing the trade-off between precision and recall in practical, real-world applications.

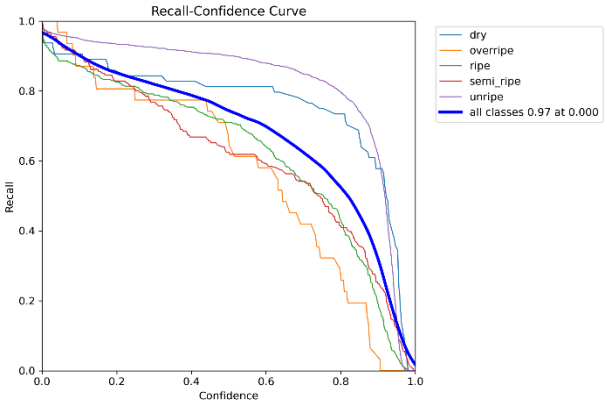

Figure 3. Recall-Confidence Curve of YOLOv8 Detection Results

Figure 3 illustrates the Recall-Confidence Curve of the YOLOv8 object detection model across five object classes: dry, overripe, ripe, semi_ripe, and unripe. This graph depicts the relationship between the confidence threshold (X-axis) and the recall value (Y-axis). In general, recall tends to decrease as the confidence threshold increases. This trend

occurs because a higher threshold leads to fewer predictions being accepted as positive, thereby reducing the model's ability to detect all actual objects present in the image.

In this curve, the unripe class again demonstrates the best performance, with consistently high and stable recall values across the full range of confidence thresholds. In contrast, the overripe class exhibits the lowest recall among all classes, particularly from a confidence level of 0.5 and above. This trend suggests that the model can effectively detect most unripe objects even at high confidence levels, but struggles to maintain recall for overripe objects as the confidence threshold becomes stricter.

The thick blue line in the graph represents the average recall across all classes, reaching a maximum value of 0.97 at a confidence threshold of 0.0. This indicates that when no confidence restriction is applied, the model is capable of detecting nearly all objects present. However, the average recall gradually declines as the threshold increases, approaching zero at a threshold of 1.0. This curve is crucial for determining an appropriate trade-off between recall and confidence threshold during deployment, ensuring a balance between maximizing correct detections and minimizing false positives in real-world applications.

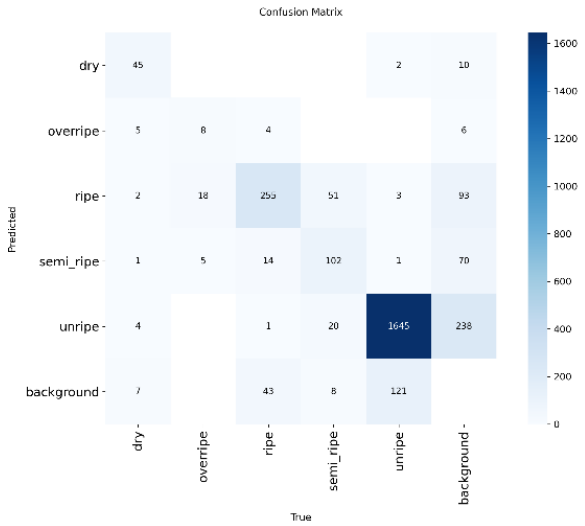

Figure 4. Confusion Matrix of Object Detection Results Using YOLOv8 Model

Figure 4 presents the object detection results obtained using the YOLOv8 model on a coffee fruit dataset consisting of six classes: dry, overripe, ripe, semi_ripe, unripe, and background. This confusion matrix represents the number of correct and incorrect predictions made by the model for each class compared to the actual labels. A higher value along the diagonal indicates better model performance in accurately classifying objects into their respective classes.

According to the results, the unripe class achieved the highest number of correct predictions, totaling 1,633 instances, indicating the model's strong capability in recognizing this class. In contrast, the overripe class showed the lowest performance, with only 13 correct predictions and a considerable number of misclassifications into other classes, particularly background, semi_ripe, and ripe. The ripe class was correctly identified in 209 instances, but it also experienced a substantial number of misclassifications into unripe and background, with 133 and 47 instances, respectively. Additionally, the semi_ripe class yielded 96 correct predictions, although it also suffered notable misclassifications into unripe and background. The dry class was correctly predicted in 51 instances, yet misclassifications into several other classes were still observed. Overall, the confusion matrix results indicate that while the YOLOv8 model performs well in classifying certain categories, such as unripe, its classification accuracy for overripe, semi_ripe, and background classes still requires improvement. Factors such as data imbalance, visual similarity between classes, and object overlap within images likely contribute to the observed misclassification patterns.

B. YOLOv8-Based Detection

The object detection process in this study was conducted using the YOLOv8 model, the latest version in the You Only Look Once (YOLO) algorithm family, offering improved accuracy and detection speed. YOLOv8 operates as a single-stage detector, performing both bounding box prediction and object classification simultaneously within a single inference step. The model was trained using a dataset of coffee fruit images consisting of five object classes: dry, overripe, ripe, semi_ripe, and unripe, as well as an additional background class to identify non-object areas.

During training, YOLOv8 demonstrated strong performance in localizing and classifying objects in the images. This is supported by evaluation metrics, where mAP50 and mAP50-95 scores consistently improved, surpassing 0.85 by the end of training, and by the steadily declining loss curves up to epoch 100. Additionally, the generated precision and recall curves showed stable and positive trends with increasing confidence thresholds, indicating the model's ability to maintain a balance between detection precision and completeness.

Overall, the object detection results using YOLOv8 in this study proved effective in recognizing and detecting coffee fruits in images. However, there remains room for improvement in classifying visually similar classes. To enhance the accuracy of the classification results, post-detection classification refinement was performed through the integration of the Vision Transformer (ViT), which will be discussed in the following section. Figure 5 illustrates an example of successful multi-class detection of coffee fruits directly from their bunches using YOLOv8.

Figure 5 illustrates an example of object detection results using the YOLOv8 model on an image of a coffee fruit bunch containing various ripeness levels. In this image, the model detects five object classes simultaneously: dry, overripe, ripe, semi_ripe, and unripe. Each detected object is marked with a color-coded bounding box according to its category, along with a confidence score indicating the model's certainty for each prediction. The detection was performed directly on the intact bunch image without prior cropping, to evaluate the

model's ability to recognize objects under natural field conditions.

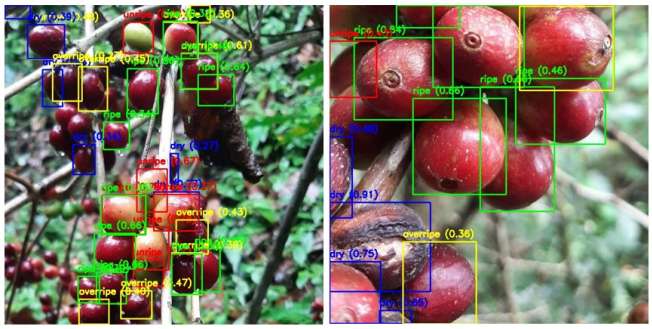

Figure 5. Example of Coffee Fruit Detection Results by YOLOv8 Directly from the Bunch with Multi-Class Labels

The results show that the model is capable of identifying most objects effectively, particularly for the ripe and dry classes, which were predominantly detected with confidence scores above 0.7. In contrast, some classes such as overripe and semi_ripe were detected with more variable confidence levels, including scores below 0.5. This suggests challenges in distinguishing visual characteristics between classes with closely related ripeness stages. Additionally, instances of duplicate or overlapping detections on the same objects were observed, highlighting the complexity of simultaneous multi-class detection within a single bunch image.

As an illustrative result, this figure demonstrates the performance of YOLOv8 in a real-world application context, where objects are naturally distributed within a frame. It also serves as an initial stage before object cropping is conducted—based on bounding boxes—for enhanced classification using the Vision Transformer (ViT) model in the subsequent phase.

C. Integration of YOLOv8 and Vision Transformer

Following the object detection process using YOLOv8, the resulting bounding boxes surrounding each object in the coffee bunch images are utilized as the basis for further classification. At this stage, each object detected by YOLOv8 is automatically cropped based on its bounding box coordinates, resulting in smaller image segments containing only a single object, free from background interference or neighboring objects. These cropped images are subsequently processed using the Vision Transformer (ViT) model to enhance the initial classification outcomes produced by YOLOv8. Figure 6 illustrates the automatic cropping of YOLOv8 bounding boxes, which are then classified using ViT.

Figure 6 illustrates examples of the automatic cropping process generated based on YOLOv8 detection bounding boxes applied to images of coffee bunches. Each cropped image represents a single, intact coffee cherry with varying levels of ripeness corresponding to its designated class. In this study, five object classes are utilized: dry, overripe, ripe, semi_ripe, and unripe. Each cropped segment reflects distinct visual characteristics of its class, including skin texture, color, and surface conditions.

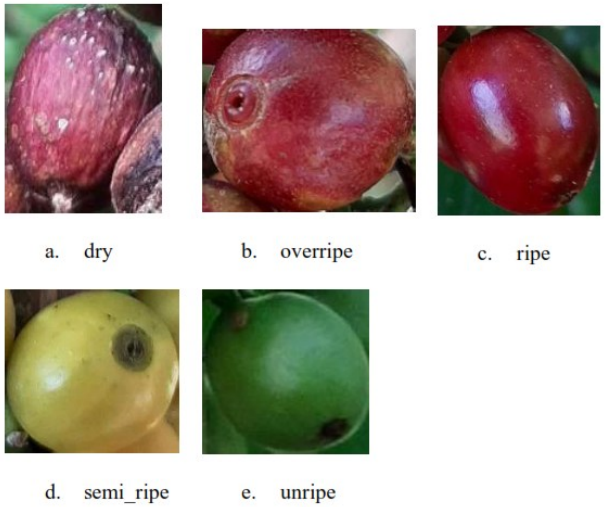

Figure 6. Results of the Automatic Cropping Process Based on YOLOv8 Detection Bounding Boxes

The cropping process is executed automatically using the detection coordinates produced by YOLOv8, without any manual intervention, resulting in a new image dataset consisting solely of isolated objects. These cropped images are subsequently employed for further classification using the Vision Transformer (ViT). The purpose of this process is to ensure that ViT focuses exclusively on the object area without background interference, thereby improving classification accuracy by eliminating surrounding visual noise. Consequently, these automatically cropped images serve not only to reinforce the initial YOLOv8 classification results but also to play a crucial role in the integration stage of the ViT-based detection and classification system. The final results of detection using the integrated YOLOv8 and Vision Transformer model are presented in Figure 7.

Figure 7 presents the final outcome of the coffee fruit detection and classification system that integrates YOLOv8 with the Vision Transformer (ViT) to enhance classification accuracy. In the initial stage, YOLOv8 performs direct multi-class detection on objects within the coffee bunch image, generating bounding boxes colored according to the initial predicted classes. These detected object segments are then cropped and further processed independently by ViT, which reclassifies each object based on its visual features. The classification results from ViT are subsequently compared to the initial YOLOv8 predictions.

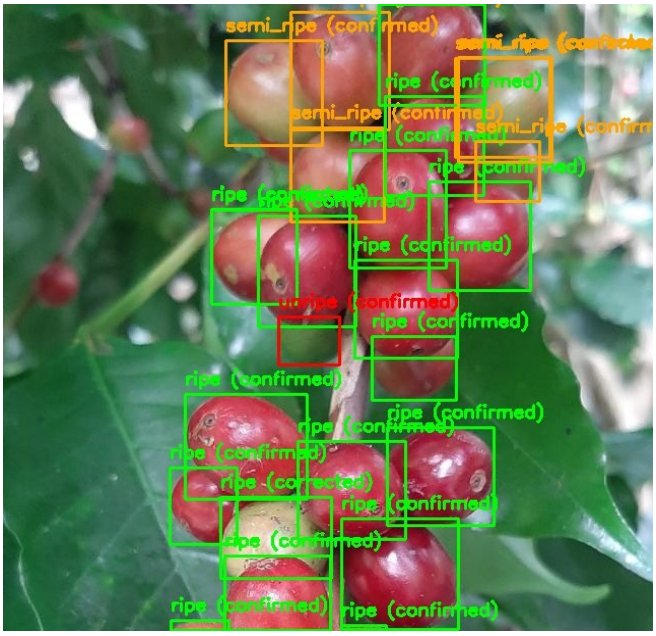

Figure 7. Example of Detection Results Using YOLOv8 + Vision Transformer

In Figure 7, each bounding box displays two pieces of information: the class label predicted by YOLOv8 and the confirmation status from ViT. The label confirmed indicates that the ViT classification matches the initial YOLOv8 prediction, whereas the label corrected signifies that ViT has revised an inaccurate YOLOv8 prediction. For instance, some objects initially classified as semi_ripe by YOLOv8 were corrected to ripe by ViT, or vice versa.

This integration has proven effective in improving final classification accuracy, particularly for visually similar classes such as ripe and semi_ripe. Moreover, the system reduces prediction errors for objects with low confidence scores in YOLOv8 by leveraging ViT as a secondary verifier based on deep visual feature analysis. This figure visually demonstrates the effectiveness of combining two deep learning-based approaches within a single object detection and image transformation classification system, offering strong potential for applications in precision agriculture.

D. Performance Comparison Between YOLOv8 and YOLOv8 + Vision Transformer

To evaluate the effectiveness of integrating the Vision Transformer (ViT) as a classification enhancer, a performance comparison was conducted between the detection and classification results obtained using YOLOv8 alone and those achieved after applying verification with ViT. This comparison focuses on key performance metrics, including per-class classification accuracy, mean Average Precision (mAP), precision, recall, and confusion matrix analysis for both scenarios.

The results indicate that YOLOv8 alone provides reasonably good detection and classification performance, achieving an mAP50 of 0.88 and average recall above 0.80 for the ripe and unripe classes. However, detection performance for the overripe and semi_ripe classes remains suboptimal, as evidenced by lower precision and recall values and relatively high misclassification rates into neighboring classes. These challenges are largely attributed to class imbalance and the high visual similarity among certain fruit maturity levels.

Upon integrating the Vision Transformer, the system exhibits significant improvements, particularly in the classification accuracy of the semi_ripe and overripe classes. ViT functions as a verification module that reclassifies cropped object segments generated by YOLOv8. When the initial classification yields a confidence score below a predefined threshold or shows disagreement with ViT results, the final prediction is adjusted according to ViT's output. This approach enhances the system's average precision by approximately 4–7% and reduces classification errors, especially those occurring between semi_ripe and ripe classes.

Furthermore, the integrated system demonstrates improved prediction consistency by reducing the occurrence of duplicate detections (overlapping detection) and misdetections of small objects with low confidence scores. Overall, the comparative analysis confirms that the combination of YOLOv8 and Vision Transformer yields superior classification performance compared to YOLOv8 alone, without significantly compromising system speed or computational efficiency. Figure 8 illustrates an example comparison of detection results using YOLOv8 versus the integrated YOLOv8 + Vision Transformer approach.

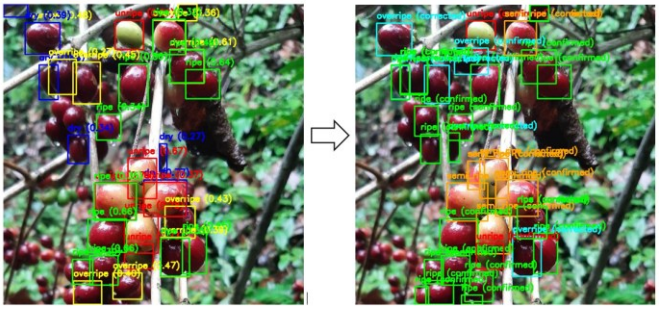

Figure 8. Comparison of Detection Results Between YOLOv8 Alone and YOLOv8 + Vision Transformer

In this study, the performance evaluation of the coffee fruit detection and classification models was conducted by comparing the results of YOLOv8 alone with those of YOLOv8 integrated with the Vision Transformer (ViT) as a classification enhancer. The evaluation was carried out under three different training epoch scenarios: 50, 100, and 150, using the same training parameters for both models, namely a learning rate of 0.0001 and a batch size of 16. The objective of this evaluation was to assess the extent of improvement in classification accuracy, detection stability, and class distribution correction introduced by incorporating ViT into the system.

The results indicate that YOLOv8 achieves stable performance as the number of epochs increases, with precision, recall, and mAP values gradually improving and

reaching optimal levels at 150 epochs. However, when YOLOv8 is combined with the Vision Transformer, the overall classification performance shows a more substantial improvement, particularly in the precision and mAP50-95 metrics. This enhancement is evident from the higher metric values observed at each epoch compared to YOLOv8 alone, with the most notable differences occurring at epochs 100 and 150.

In addition to numerical metric-based evaluation, this study also presents a confusion matrix-based analysis to compare the distribution of misclassification errors across classes for both models. The confusion matrix for YOLOv8 reveals a relatively high rate of misclassification among the ripe, semi_ripe, and overripe classes. Several ripe objects were misclassified as semi_ripe, or vice versa, due to visual similarities between these categories. After integrating the Vision Transformer, the confusion matrix demonstrates significant improvements. The number of misclassifications between similar classes was reduced, particularly for semi_ripe and overripe, while the number of correct predictions increased across nearly all classes. The confusion matrix results for the 100-epoch scenario are presented in Figure 9.

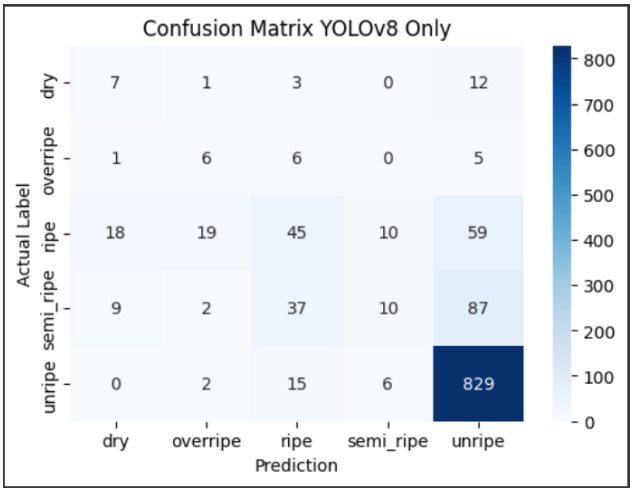

Figure 9. Confusion Matrix for YOLOv8 Only

The confusion matrix presented in Figure 9 illustrates the classification performance of the YOLOv8 model on cropped coffee fruit objects extracted from the original cluster images. Each detected object was automatically cropped using the bounding box coordinates generated by YOLOv8, and the predicted labels were then compared with their corresponding ground truth labels to analyze the class-wise distribution of predictions.

The results indicate that the unripe class achieved the highest number of correct predictions, with 829 samples correctly classified. However, several other classes exhibited substantial misclassification rates. For instance, within the ripe class, only 45 samples were accurately classified, while 59 were incorrectly predicted as unripe, 18 as dry, 19 as overripe, and 10 as semi_ripe. Similarly, for the semi_ripe class, only 37 samples were correctly classified, whereas 87 were misclassified as unripe, and 37 others were dispersed among various other classes.

These findings highlight the challenges YOLOv8 faces in distinguishing between coffee fruit categories with visually similar characteristics. Consequently, the integration of the Vision Transformer was proposed to re-verify the YOLOv8 cropping results, thereby enhancing the final classification accuracy through deeper visual feature analysis.

The confusion matrix presented in Figure 10 illustrates the classification performance of the integrated system combining YOLOv8 with Vision Transformer (ViT), using cropped samples of coffee fruit objects obtained automatically from YOLOv8 detections. Each detected object was cropped based on the bounding box coordinates provided by YOLOv8 and subsequently re-classified by the Vision Transformer to verify and refine the final prediction.

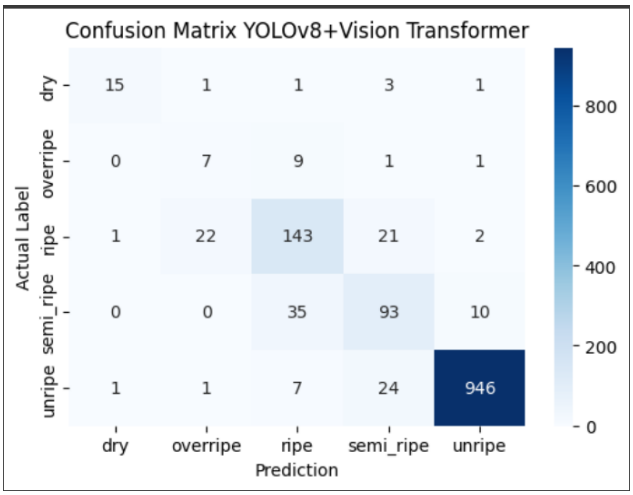

Figure 10. Confusion Matrix of YOLOv8 + Vision Transformer

The results demonstrate a significant improvement compared to using YOLOv8 alone. For the unripe class, the number of correct predictions increased to 946 samples, up from 829 previously. Similarly, in the ripe class, correctly classified instances rose to 143 from just 45. Misclassification rates across classes also decreased notably. For example, the number of ripe fruits incorrectly classified as unripe dropped from 59 to only 22. The semi_ripe class also experienced improvement, with 93 correctly classified objects, accompanied by a substantial reduction in misclassifications into the unripe category. These results confirm the effectiveness of Vision Transformer in enhancing the initial YOLOv8 classification, particularly for classes with high visual similarity. Thus, the integration of these two models significantly improves the accuracy and precision of the multi-class classification system for coffee fruit.

This study provides strong evidence that Vision Transformer functions effectively as a post-classification enhancer, refining YOLOv8's initial predictions through re-verification of the cropped objects. Notably, this integration improves final classification accuracy without altering the

core training parameters, relying instead on post-prediction enhancement. Overall, the combined YOLOv8 and Vision Transformer system delivers superior and more stable multi-class classification performance compared to the standalone YOLOv8 approach. The complete summary of the experimental results can be found in Table 1.

TABLE 1.
SUMMARY OF OVERALL EXPERIMENTAL RESULTS

Model	Epoch	Akurasi (%)	Precision (%)	Recall (%)
YOLOv8	50	**69.02%**	70.91%	69.02%
YOLOv8	100	**75.44%**	70.65%	75.44%
YOLOv8	150	**70.04%**	68.02%	70.04%
YOLOv8 + ViT	50	**89.02%**	88.80%	89.02%
YOLOv8 + ViT	100	**89.52%**	90.43%	89.52%
YOLOv8 + ViT	150	**88.68%**	75.53%	65.62%

Table 1 presents the evaluation of detection and classification performance for coffee fruit in this study, comparing two model schemes: YOLOv8 alone and an integrated model combining YOLOv8 with Vision Transformer (ViT) as a classification enhancer. The training process was conducted using consistent parameters for both models, with a learning rate of 0.0001 and a batch size of 16. Performance testing was carried out across three epoch variations—50, 100, and 150—to assess the model's stability with increasing training iterations.

Based on the results presented in Table 1, the YOLOv8-only model exhibited a gradual improvement in accuracy, reaching 75.44% at epoch 100, compared to 69.02% at epoch 50. However, its performance declined at epoch 150, with accuracy falling to 70.04%. The precision and recall metrics followed a similar trend, with the highest precision recorded at epoch 50 (70.91%), which then decreased at epochs 100 and 150. This trend suggests that YOLOv8 may suffer from overfitting or performance instability when the number of training epochs is increased excessively.

In contrast, the integration of YOLOv8 with Vision Transformer demonstrated significantly more stable and superior performance. At epoch 50, the integrated model achieved an accuracy of 89.02%, with a precision of 88.80% and a recall of 89.02%. Performance continued to improve at epoch 100, reaching the highest accuracy of 89.52% and precision of 90.43%. Although a slight drop in accuracy to 88.68% was observed at epoch 150, the precision and recall values remained within a high-performance range. Specifically, while precision fluctuated to 75.53% at epoch 150, recall experienced a more pronounced decrease to 65.62%. This decline is likely attributed to overfitting tendencies at higher epoch levels. Nevertheless, the overall performance of the integrated model confirms the effectiveness of Vision Transformer in enhancing classification accuracy and consistency compared to YOLOv8 alone.

The performance differences indicate that the Vision Transformer plays a crucial role in strengthening the classification outcomes, particularly in distinguishing coffee fruit classes with visually similar characteristics. This integration maintains high system accuracy even with increasing training epochs and improves detection reliability during post-prediction stages—without requiring adjustments to the fundamental training parameters.

E. Discussion

Based on the results of this study, the integration of YOLOv8 with Vision Transformer (ViT) significantly enhances the performance of the multi-class detection and classification system for coffee fruits compared to using YOLOv8 alone. The automatic cropping process of coffee fruit objects using YOLOv8 plays a critical role in the system, as the entire evaluation is conducted based on these cropped objects. The resulting confusion matrices indicate that YOLOv8 still struggles to distinguish between visually similar classes of coffee fruits, particularly among ripe, semi_ripe, and unripe categories. This is evident from the high number of misclassifications within these classes.

The implementation of Vision Transformer as a post-prediction verification module effectively reduces classification errors. After YOLOv8 generates the bounding boxes, ViT performs a more precise re-classification of each cropped object. This improvement is reflected in the increased number of correct predictions for the ripe, semi_ripe, and unripe classes in the confusion matrix resulting from the YOLOv8+ViT integration. Additionally, improvements are also observed across overall evaluation metrics. Without ViT, YOLOv8 achieves a maximum accuracy of only 75.44% at epoch 100, whereas the integrated model reaches a significantly higher accuracy of 89.52% at the same epoch.

Beyond accuracy enhancement, the integration of ViT also positively impacts precision and recall metrics. The highest precision is achieved by YOLOv8+ViT at epoch 100 with a score of 90.43%, which is substantially higher than the precision of YOLOv8 alone at 70.65%. Similarly, the recall value remains above 89% up to epoch 100 before slightly decreasing at epoch 150, which is likely attributed to overfitting as a result of excessive training iterations. This indicates the importance of optimizing the number of epochs to maintain model stability.

From the visual detection results, both the output of YOLOv8 alone and the version verified by ViT demonstrate the model's capability in identifying coffee fruit objects within bunches. However, misclassification still occurs among classes with similar color and texture. The presence of the Vision Transformer helps correct these weaknesses, thus improving the final detection and classification results by providing prediction confirmation within each object label. This integration proves effective for multi-class image classification systems in agricultural applications, especially where inter-class visual similarity is high.

Overall, this study demonstrates that the Vision Transformer serves as a powerful enhancer for YOLOv8-based object detection classification, without requiring any modification to the core training parameters. The combination of both methods yields more optimal results than using either method independently and can be applied to various related

studies involving multi-class classification of small objects in complex images.

IV. CONCLUSION

This study successfully developed a multi-class coffee fruit detection and classification system by integrating YOLOv8 with the Vision Transformer (ViT) as a classification enhancement module. Based on the evaluation results, YOLOv8 demonstrated satisfactory object detection capabilities; however, it still encountered challenges in distinguishing between classes with similar visual characteristics, such as ripe, semi_ripe, and unripe coffee fruits. This limitation was evident from the confusion matrix and evaluation metrics, which indicated suboptimal precision and recall values, along with noticeable misclassifications between classes.

The integration of the Vision Transformer as a post-prediction verification stage significantly improved system performance. The evaluation results revealed that the YOLOv8+ViT configuration achieved an accuracy of up to 89.52%, with a precision of 90.43% and a recall of 89.52% at epoch 100—substantially higher than the YOLOv8-only model. Furthermore, the number of misclassifications across visually similar classes was significantly reduced.

In conclusion, the application of the Vision Transformer as a classification enhancer in a YOLOv8-based object detection system is effective in improving overall accuracy, precision, and recall. This integration shows strong potential for implementation in multi-class classification systems in the agricultural domain or for small object recognition in visually complex conditions. However, a limitation of this study is that the proposed method has not yet been evaluated for real-time deployment in field conditions, which remains a challenge for practical implementation. Future work may explore a wider range of training parameters or incorporate ensemble methods to further enhance the performance of deep learning-based classification systems.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to the Ministry of Education, Culture, Research, and Technology (Kemdikbudristek) for providing financial support through the Penelitian Dosen Pemula (Beginner Lecturer Research) scheme in 2025. We also highly appreciate the collaboration and contributions of fellow researchers and partner institutions that have supported the successful implementation of this study. With such support, this research was able to proceed smoothly and produce valuable findings for the development of computer vision-based technology for detecting coffee fruit ripeness.

REFERENCES

[1] A. Subki and B. Imran, "Implementasi Deep Learning Menggunakan CNN dengan Arsitektur Alexnet Untuk Klasifikasi dan Identifikasi Jenis Kopi Khas Lombok Ahmad," *Explore*, vol. 14, no. 2, pp. 135–140, 2024.

[2] N. Pradita, Hayati, Suwardji, Muktasam, and Mulyati, "Analisis Keberlanjutan Dimensi Ekologi Kopi Arabika di Lahan Kering Desa Sajang Kecamatan Sembalun Kabupaten Lombok Timur," *Agroteksos*, vol. 34, no. 2, pp. 383–391, 2024.

[3] L. Y. K. Chandra, B. I. Linggarweni, and S. Novida, "Analisis Pendapatan Usaha Kopi Bubuk Arabika di Desa Sajang Kecamatan Sembalun Kabupaten Lombok Timur," *J. Ekon. dan Bisnis*, vol. 3, no. 2, pp. 148–155, 2023, doi: 10.56145/jurnalekonomidanbisnis.v3i2.71.

[4] T. C. Pham, V. D. Nguyen, C. H. Le, M. Packianather, and V. D. Hoang, "Artificial intelligence-based solutions for coffee leaf disease classification," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1278, no. 1, 2023, doi: 10.1088/1755-1315/1278/1/012004.

[5] E. Elbasi *et al.*, "Artificial Intelligence Technology in the Agricultural Sector: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 171–202, 2022.

[6] B. Ye, R. Xue, and H. Xu, "ASD-YOLO: a lightweight network for coffee fruit ripening detection in complex scenarios," *Front. Plant Sci.*, vol. 16, no. February, pp. 1–13, 2025, doi: 10.3389/fpls.2025.1484784.

[7] H. C. Bazame, J. P. Molin, D. Althoff, and M. Martello, "Detection of coffee fruits on tree branches using computer vision," *Sci. Agric.*, vol. 80, no. October, 2022, doi: 10.1590/1678-992X-2022-0064.

[8] S. Velásquez, A. P. Franco, N. Peña, J. C. Bohórquez, and N. Gutiérrez, "Classification of the maturity stage of coffee cherries using comparative feature and machine learning," *Coffee Sci.*, vol. 16, no. March, p. 1, 2021, doi: 10.25186/.v16i.1710.

[9] M. N. Izza and G. P. Kusuma, "Image Classification of Green Arabica Coffee Using Transformer-Based Architecture," *Int. J. Eng. Trends Technol.*, vol. 72, no. 6, pp. 304–314, 2024, doi: 10.14445/22315381/IJETT-V72I6P128.

[10] M. García, J. E. Candelo-Becerra, and F. E. Hoyos, "Quality and defect inspection of green coffee beans using a computer vision system," *Appl. Sci.*, vol. 9, no. 19, 2019, doi: 10.3390/app9194195.

[11] H. L. Gope, H. Fukai, F. M. Ruhad, and S. Barman, "Comparative analysis of YOLO models for green coffee bean detection and defect classification," *Sci. Rep.*, vol. 14, no. 1, pp. 1–16, 2024, doi: 10.1038/s41598-024-78598-7.

[12] A. Rincon-Jimenez *et al.*, "Ripeness stage characterization of coffee fruits (coffea arabica L. var. Castillo) applying chromaticity maps obtained from digital images," in *Materials Today: Proceedings*, Elsevier Ltd., 2021, pp. 1271–1278. doi: 10.1016/j.matpr.2020.11.264.

[13] A. Michael and M. Garonga, "Classification model of 'Toraja' arabica coffee fruit ripeness levels using convolution neural network approach," *Ilk. J. Ilm.*, vol. 13, no. 3, pp. 226–234, 2021, doi: 10.33096/ilkom.v13i3.861.226-234.

[14] A. G. Costa, D. A. G. De Sousa, J. L. Paes, J. P. B. Cunha, and M. V. M. De Oliveira, "Classification of robusta coffee fruits at different maturation stages using colorimetric characteristics" *Eng. Agrícola*, vol. 4430, no. 4, pp. 518–525, 2020, [Online]. Available: https://doi.org/10.1590/1809-4430-Eng.Agric.v40n4p518-525/2020

[15] B. Xiao, M. Nguyen, and W. Q. Yan, "Fruit ripeness identification using YOLOv8 model," *Multimed. Tools Appl.*, vol. 83, no. 9, pp. 28039–28056, 2024, doi: 10.1007/s11042-023-16570-9.