Vol. 15, No. 3 (2026): 1164 - 1179 https://dx.doi.org/10.23960/jtepl.v15i3.1164-1179

JURNAL TEKNIK PERTANIAN LAMPUNG

ISSN 2302-559X (print) / 2549-0818 (online)
Journal homepage : https://jurnal.fp.unila.ac.id/index.php/JTP

Non-Destructive Detection of Coffee Bean Defects using Machine Vision and the YOLOv11 Algorithm

Hary Kurniawan[1,✉], Ince Siti Wardatullatifah S.[1], Hanifah Ayu[1], Surya Abdul Muttalib[1], Sukmawaty[1], Ansar[1], Rahmat Sabani[1], Murad[1]

[1] Department Agriculture Engineering, Faculty of Food Technology and Agroindustry, University of Mataram, Nusa Tenggara Barat, INDONESIA.

Article History:

Received : 9 February 2026
Revised : 9 March 2026
Accepted : 13 March 2026

Keywords:

Coffee beans,
Defect detection,
Machine vision,
YOLOv11.

Corresponding Author:
✉ harykurniawan@unram.ac.id
(Hary Kurniawan)

ABSTRACT

Advances in machine vision and deep learning offer a promising solution for automated, non-destructive quality assessment for high-quality coffee. This study evaluated the performance of five YOLOv11 variants (n, s, m, l, and x) for real-time detection of defective coffee beans and identified the most suitable model in terms of detection accuracy and computational efficiency. A conveyor-based machine vision system was developed to acquire top-view images of Robusta coffee beans under controlled illumination. A dataset of 3,500 images was prepared, comprising 3,000 annotated images for training and validation (80:20) and 500 images reserved for blind testing. All defective beans were grouped into a single defect class, and the YOLOv11 variants were evaluated using precision, recall, F1-score, mean Average Precision (mAP) at IoU thresholds of 0.5 and 0.5:0.95, and inference time. All YOLOv11 variants achieved high detection performance, with mAP@0.5 values exceeding 0.98. YOLOv11s showed the best overall balance, achieving the highest recall (0.954), mAP@0.5:0.95 (0.689), and F1-score (0.959), while maintaining low inference time and a compact model size. Larger variants, such as YOLOv11x, achieved slightly higher mAP@0.5 but required substantially greater computational resources, whereas YOLOv11n provided faster inference but lower robustness under stricter localization criteria. Blind testing revealed a performance gap relative to validation results, highlighting remaining challenges in model generalization. Overall, the results confirm the effectiveness of YOLOv11 for coffee bean defect detection and identify YOLOv11s as the most suitable variant for real-time inspection within the defined experimental scope.

1. INTRODUCTION

Coffee is one of the most economically valuable agricultural commodities worldwide, with more than two billion cups consumed daily, making it one of the most widely traded beverages globally (Al-Ghamdi *et al*., 2024; Gosalvitr *et al*., 2024). The growing demand for high-quality coffee requires proper bean selection, as bean quality strongly influences market price, consumer acceptance, and international competitiveness (Al-Mahish *et al*., 2024). Physical quality assessment generally relies on visual attributes such as color, morphology, shape, and size, all of which contribute to sensory properties and storage stability (Chandu *et al*., 2025). During pre-trading stages, quality variation is strongly affected by the presence of visual or structural defects that can alter the sensory profile of the final product. Therefore, quality assurance practices are essential for maintaining consistency and competitiveness in the global coffee market (Anastácio *et al*., 2023; Matias *et al*., 2025).

Specific defects, such as black beans, broken beans, insect damage, and sour beans, have been shown to reduce both sensory quality and economic value; in green coffee trading, such defects may even lead to lot rejection by international buyers (Ji *et al*., 2024; Thai *et al*., 2024). Traditionally, quality inspection is still widely performed manually through visual sorting by workers to identify defects and separate acceptable beans (S.-Y. Chen *et al*., 2022; Wang *et al*., 2021). However, this manual procedure is subjective, slow, and highly dependent on worker fatigue, lighting conditions, and experience, resulting in inconsistent outcomes (Arwatchananukul *et al*., 2024). Mechanical sorting methods based on size, weight, or shape are relatively faster, but they often fail to detect subtle visual defects such as discoloration or minor damage and may even damage beans because of the invasive nature of the sorting process (García *et al*., 2019). These limitations highlight the need for advanced detection systems that are more objective, rapid, and efficient for identifying coffee defects, thereby supporting improvements in the quality and market value of green coffee (Ji *et al*., 2024).

Advances in machine vision combined with artificial intelligence (AI) have created substantial opportunities to improve non-destructive food quality inspection, including that of coffee beans (Lin *et al*., 2023; Wang *et al*., 2021). Machine vision systems enable the automatic extraction of visual features such as color, shape, and texture, allowing defective beans to be identified objectively, consistently, and efficiently without manual intervention (Hemamalini *et al*., 2022; Shen *et al*., 2024). Compared with mechanical or conventional visual inspection methods, this approach is not only faster but also capable of handling large volumes of data with high accuracy, making it highly relevant for industrial-scale production lines (Lin *et al*., 2023; Ali *et al*., 2021). Among deep learning algorithms, You Only Look Once (YOLO) has emerged as one of the most widely used object detection frameworks because it performs classification and localization simultaneously in real time with high accuracy (Alif *et al*., 2024). Comprehensive overviews of recent developments, variants, and applications of the YOLO framework in object detection can be found in the review studies of Badgujar *et al*. (2024) and Murat & Kiran (2025).

Recent studies have demonstrated the effectiveness of computer vision and deep learning, particularly the YOLO family, for automating coffee bean defect detection across Arabica, Robusta, and Liberica varieties, thereby addressing the inefficiency and subjectivity of manual inspection (Amadea *et al*., 2024; Chang *et al*., 2024). Implementations of YOLOv7 and YOLOv8 have shown strong performance in identifying multiple defect categories, including broken, sour, immature, insect-infested, and moldy beans, even in images containing multiple overlapping coffee beans (Gope *et al*., 2024; Liang *et al*., 2023). In particular, YOLOv8n has achieved high accuracy, precision, recall, and mean average precision (mAP), while the integration of Grad-CAM has improved model interpretability by highlighting the image regions that contribute most strongly to predictions, thereby increasing its reliability for quality assessment tasks (Adiwijaya *et al*., 2024; Oncu, 2025). In parallel, multi-model frameworks combining YOLOv7 with convolutional neural networks (CNNs) have further improved classification performance, reaching an accuracy of up to 98.97%, which highlights the complementary strengths of detection and classification architectures in defect analysis (Liang *et al*., 2023). Comparative evaluations of YOLOv8 and YOLOv10 have also indicated that both models are suitable for real-time applications, with YOLOv8n providing slightly higher accuracy and YOLOv10n offering marginally faster inference, thus presenting scalable solutions for industrial deployment (Adiwijaya *et al*., 2024). Beyond the YOLO family, defect detection has also been successfully applied to different coffee varieties using alternative deep learning architectures. For Arabica beans, models such as LSKNet and modified VGG16 have demonstrated strong detection performance, with LSKNet-S achieving an mAP of 0.879 and outperforming several baseline variants (Ardian *et al*., 2024; Samudra & Rachmawati, 2025). Similarly, for Robusta coffee beans, YOLOv5-based systems have attained an overall accuracy of 95.11%, offering an efficient and objective alternative to manual inspection for both producers and consumers (Luis *et al*., 2022).

Based on these advances and the remaining challenges in automated coffee bean defect detection, the present study proposes a machine vision-based approach for the non-destructive, rapid, and objective inspection of defective coffee beans using the YOLOv11 algorithm. Rather than functioning solely as a comparison of model variants, this study contributes by integrating a controlled image acquisition system, a dedicated coffee bean defect image dataset, and a systematic evaluation of five YOLOv11 variants to identify the most suitable model for practical inspection purposes. Specifically, the study focuses on: (i) developing a machine vision framework for coffee bean defect detection under controlled imaging conditions; (ii) constructing a comprehensive image dataset covering major coffee bean defect conditions; and (iii) evaluating the detection performance, robustness, and efficiency of YOLOv11 variants using standard metrics, including precision, recall, F1-score, and mean average precision (mAP). Through this integrated framework, the study aims to provide a more objective and efficient alternative to conventional manual coffee quality inspection and to establish a technical basis for future real-time automated inspection systems.

2. MATERIALS AND METHODS

2.1. YOLOv11

YOLOv11 represents a recent development within the YOLO family of object detection models, introduced in late 2024 as a single-stage, real-time object detection framework that extends earlier YOLO versions through substantial architectural refinements in its backbone and neck design (Khanam & Hussain, 2024). YOLOv11 is available in five model variants, YOLOv11n, YOLOv11s, YOLOv11m, YOLOv11l, and YOLOv11x which share the same overall architecture but differ in model scale and capacity (Khanam & Hussain, 2024; Rasheed & Zarkoosh, 2025). The lighter variants prioritize lower computational cost and faster inference, whereas the larger variants offer greater representational power at the expense of increased model size, training time, and latency. This scale-based variation provides the technical basis for comparing detection accuracy and computational efficiency in the present study (Kishor, 2024). Compared with its predecessors, namely YOLOv8 to YOLOv10, YOLOv11 places greater emphasis on efficient multi-scale feature extraction and attention-aware feature fusion to improve detection accuracy while preserving real-time performance (Chen *et al*., 2025; Hidayatullah *et al*., 2025). These improvements are realized through redesigned components such as the C3k2 block, Fast Spatial Pyramid Pooling (SPPF), and parallel spatial attention mechanisms, which collectively enhance contextual representation and feature discrimination across different spatial resolutions (Khanam & Hussain, 2024; Li *et al*., 2025).

As illustrated in Figure 1, the overall YOLOv11 architecture follows a backbone–neck–head paradigm, in which the optimized backbone extracts hierarchical features, the enhanced neck performs multi-scale feature aggregation, and the detection head outputs bounding boxes and class probabilities in an end-to-end manner (Chen *et al*., 2025; Zhao & Zhu, 2025). These architectural modules effectively expand the receptive field while preserving fine-grained spatial details, thereby improving the localization and classification of small, dense, or partially occluded objects under

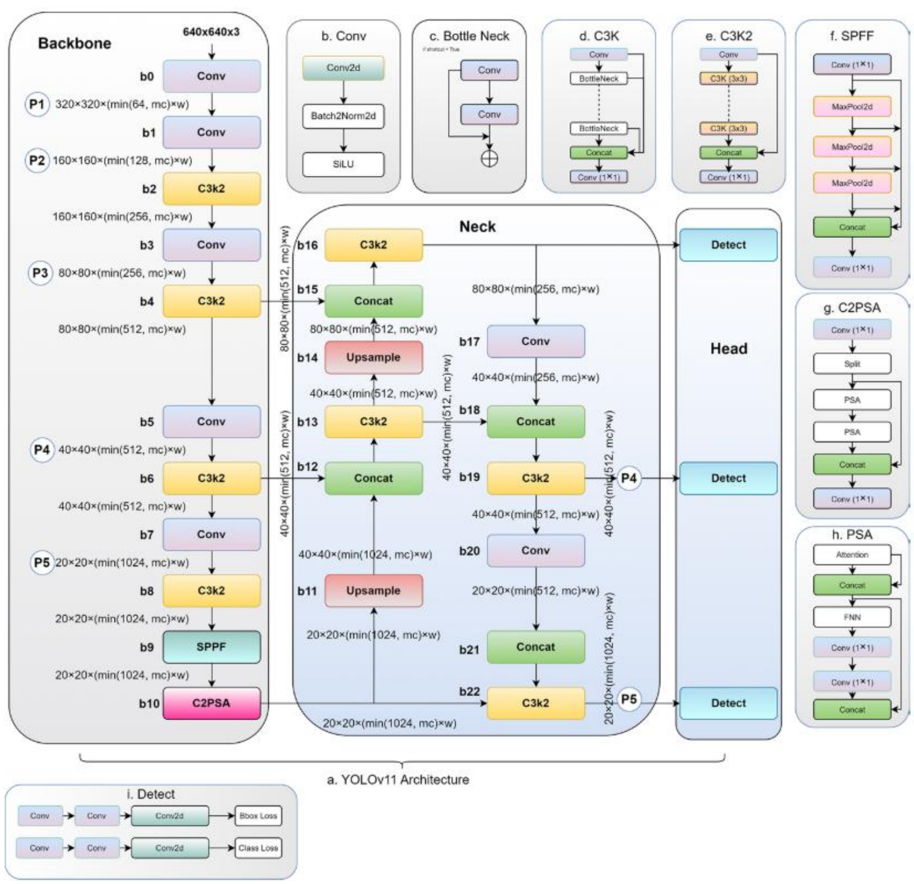


Figure 1. YOLOv11 architecture (Rasheed & Zarkoosh, 2025)

complex visual conditions (Khanam & Hussain, 2024; Zhao & Zhu, 2025). Furthermore, YOLOv11 is released in multiple model-size variants, ranging from lightweight configurations to higher-capacity models, enabling flexible trade-offs between accuracy, inference speed, and computational cost for deployment on both edge devices and server-level systems (Hidayatullah *et al*., 2025; Li *et al*., 2025). Recent benchmark evaluations across diverse application domains, including industrial inspection, agriculture, remote sensing, and medical imaging, have shown that YOLOv11 consistently achieves higher mean average precision (mAP) than earlier YOLO generations while maintaining competitive inference latency, confirming its suitability for real-time automated visual inspection tasks (Al Husaini *et al*., 2025; He *et al*., 2025; Khanam & Hussain, 2024).

2.2. Sample Preparation

The samples used in this study consisted of Robusta coffee beans obtained directly from a local coffee producer, IKM Mentari, located in Rempek Darussalam Village, North Lombok, Indonesia. The beans were supplied in two primary categories, namely defect-free beans and defective beans, to represent typical quality variation encountered in postharvest coffee processing. Prior to the experiment, the moisture content of the coffee beans was measured using a digital moisture meter and recorded at 11.63%, which falls within the recommended range for safe postharvest handling and storage of green coffee beans. This helped minimize the influence of excess moisture on visual appearance and surface characteristics.

To ensure adequate representation and statistical balance, 1 kg of defect-free beans and 1 kg of defective beans were prepared for analysis. The defective samples included common defect types frequently reported in coffee quality grading, namely black beans, broken beans, sour beans, and insect-damaged beans. These defect categories were selected based on their prevalence in Robusta coffee production and their visual distinguishability, which is important for image-based defect detection. Prior to image acquisition, all coffee beans were manually inspected and cleaned to remove dust, foreign materials, and surface residues that could interfere with image clarity or introduce noise during feature extraction.

Following preparation, the samples were stored in airtight containers at room temperature under stable environmental conditions to prevent moisture fluctuation and physical degradation before imaging. This controlled storage helped maintain a consistent sample appearance throughout the data acquisition process. A visual comparison between defect-free and defective coffee beans used in this study is presented in Figure 2, illustrating representative samples of each category and highlighting the morphological differences relevant to defect detection.

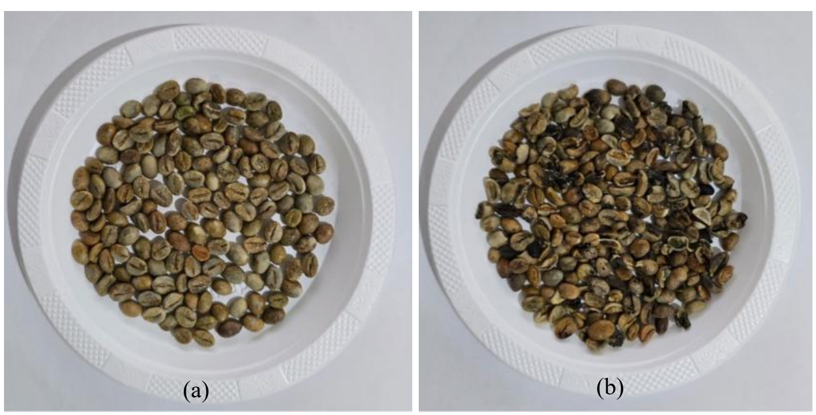


Figure 2. Comparison between (a) defect-free coffee beans and (b) defective coffee beans

2.3. Image acquisition

The defect inspection system comprised three main components: a camera, lighting units, and a conveyor mechanism, as shown in Figure 3. Image capture was performed using a 12 MP CMOS camera (F/# 2.0, focal length 4.8 mm, sensor model MA1080(5B3)CSP-481/4.5 “SAMSUNG”, resolution 1280 × 1080) mounted in a fixed overhead position 20 cm above the conveyor to obtain top-view images of the samples. The camera operated in automatic exposure mode at a maximum frame rate of 30 frame per second (fps). Illumination was provided by two 4 W LED bar lights (6500 K), installed 23.5 cm above the conveyor and positioned side by side with a spacing of 21.5 cm. Both lights were angled at 45° to provide uniform illumination and minimize shadows and reflections. The conveyor unit measured 150 cm in length, 25 cm in belt width, and 70 cm in height, and was powered by a motor equipped with a rotary speed controller ranging from 0 to 100. For this study, the dial was set to 20, corresponding to a belt speed of approximately 11 cm/s, to ensure stable transport of coffee beans during image capture.

Coffee beans were manually placed on the conveyor belt and arranged in a single layer to avoid overlap, as illustrated in **Error! Reference source not found.**a. Image acquisition and subsequent processing were conducted using a Windows 10 computer equipped with an Intel Core i7 processor, 32 GB of RAM, and an NVIDIA RTX 3060 GPU. The defect detection workflow was implemented in Python using the OpenCV and PyTorch libraries. Prior to model analysis, the coffee beans were manually inspected and cleaned to remove dust, foreign materials, and surface residues, and were then arranged in a single layer on the conveyor to prevent overlap during image capture. Images were acquired under controlled illumination and subsequently prepared for dataset construction and model development. To simulate realistic sorting conditions, defect-free and defective beans were mixed during data collection. In total, 3,500 images were captured at a resolution of 560 × 480 pixels and stored in JPG format for subsequent processing and analysis.

Description:
1. Computer;
2. Conveyor unit;
3. Camera;
4. Light source;
5. Motor;
6. Dark chamber box

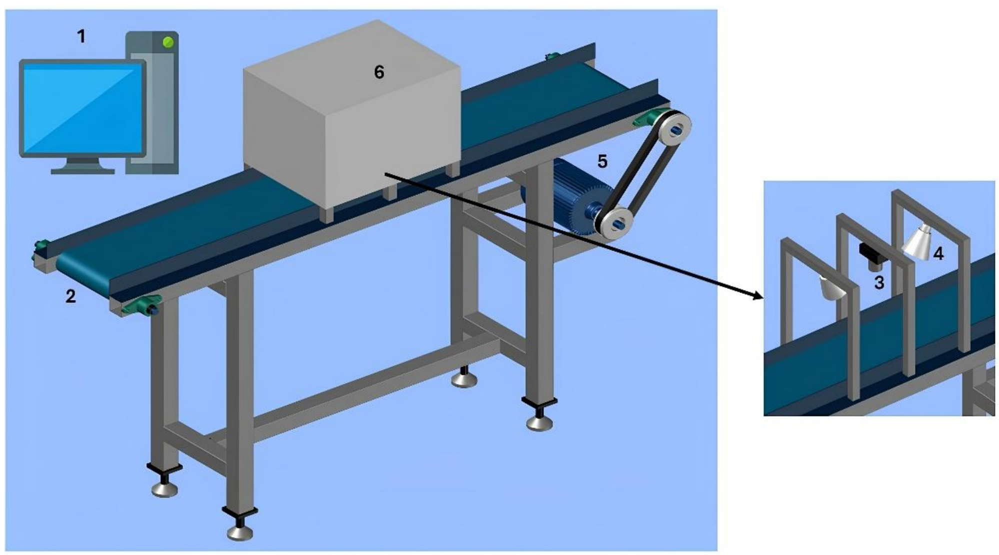


Figure 3. Developed coffee bean defect detection system

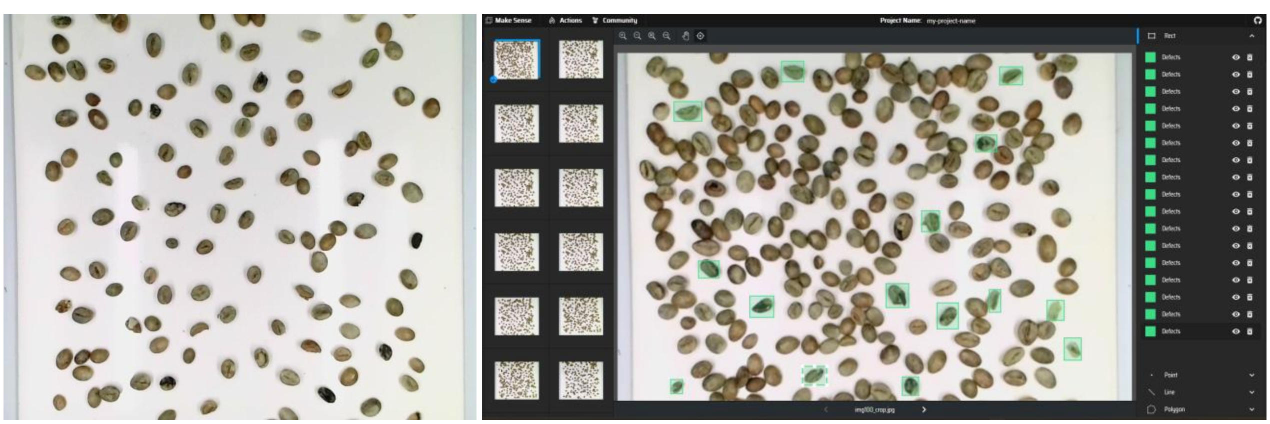


Figure 4. (a) Sample image of coffee beans mixed with defective beans used for dataset construction, (b) Image labelling

2.4. Dataset Production

A total of 3,500 images of Robusta coffee beans were collected using the imaging system described above. Of these, 3,000 images were manually annotated using *makesense.ai* (www.makesense.ai), an open-source web-based image annotation tool. The labeling process involved assigning bounding boxes to individual defective coffee beans and annotating them under a single class labeled “Dfct”, as illustrated in **Error! Reference source not found.**b. This approach simplified the detection task by grouping all defective beans into one category while ensuring accurate ground-truth data for model training and validation. Each annotated image automatically generated a corresponding .txt label file in YOLO-compatible format, containing class identifiers and normalized bounding-box coordinates. The annotated dataset was divided into training and validation subsets using an 80:20 ratio, resulting in 2,400 images for training and 600 images for validation. The training set was used to optimize the YOLOv11 models, whereas the validation set was used to monitor model performance during training and reduce the risk of overfitting. In addition, the remaining 500 images were reserved as an independent testing dataset for evaluation under unseen conditions.

The training process was conducted on a workstation equipped with the hardware specifications summarized in Table 1, while the training hyperparameters, including batch size, learning rate, optimizer, and number of epochs, are presented in Table 2. Following training, the YOLOv11 variants (n, s, m, l, and x) were systematically evaluated using key performance metrics, including precision, recall, F1-score, and mean Average Precision (mAP) at Intersection over Union (IoU) thresholds of 0.5 and 0.5:0.95. The model that provided the best trade-off between detection accuracy and computational efficiency was selected for integration into the defect inspection system.

Table 1. Hardware specifications used for model training

Configuration	Description
Programming language	Python 3.10
Operating system	Windows 10
Deep learning framework	PyTorch
GPU	NVIDIA GeForce RTX 5060Ti 16GB
CPU	13th Gen Intel(R) Core(TM)i5-13400F
Acceleration environment	CUDA 10.2 version

Table 2. Parameters used during model training

Training parameter	Description
Epochs	100
Batch-size	16
Img-size	640 × 640
Initial learning rate	0.01
Optimization algorithm	SGD
Momentum	0.937
Weight decay	0.0005

To further examine its practical applicability, the selected model was tested in real time using coffee bean samples deliberately mixed with defective beans to simulate realistic sorting conditions. This evaluation provided insight into both detection accuracy and operational reliability. The overall research workflow, from image acquisition to model evaluation and real-time testing, is summarized in Figure .

2.5. Metrics Evaluation

The performance of the YOLOv11 models in this study was evaluated using four widely recognized object detection metrics: precision, recall, F1-score, and mean Average Precision (mAP). Precision is defined as the proportion of correctly predicted positive instances (true positives) relative to the total number of predicted positive instances (true positives and false positives). A high precision value indicates that the model produces fewer false positives, which is particularly important in applications where misclassification may lead to unnecessary actions or interventions (Bumbálek *et al*., 2025).

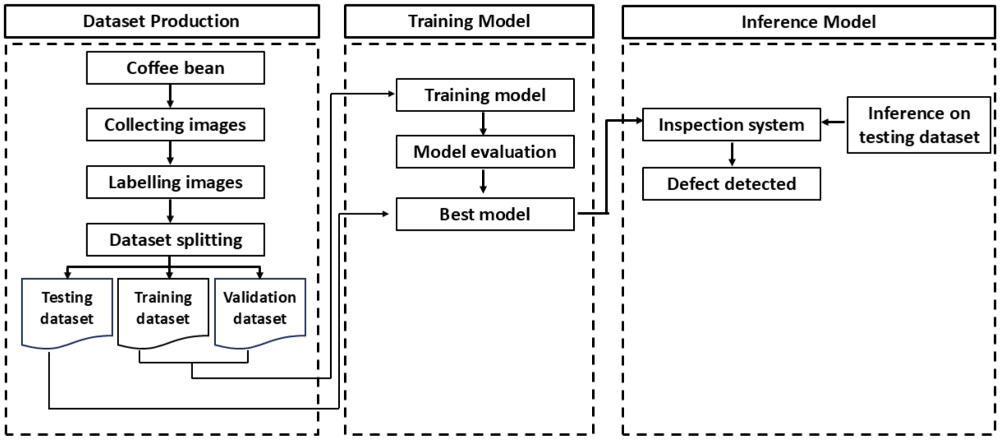


Figure 5. Research methodology workflow

$$Precision\ (P) = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{1}$$

Recall measures the proportion of actual positive instances that are correctly detected by the model, reflecting its ability to identify all relevant objects in the dataset. A higher recall value indicates that the model is more effective in minimizing missed detections (Katsamenis *et al*., 2023).

$$Recall\ (R) = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

Although precision and recall provide complementary perspectives, they are often interpreted together to obtain a balanced assessment. The F1-score, defined as the harmonic mean of precision and recall, serves this purpose by combining both metrics into a single representative value. This is particularly useful in object detection tasks, where a trade-off often exists between minimizing false positives and false negatives (Xia *et al*., 2023).

$$F1 - Score = 2 \times \frac{P \times R}{P + R} \tag{3}$$

In addition, mean Average Precision (mAP) is one of the most widely used benchmark metrics for evaluating object detection models. The mAP metric calculates the mean of average precision values across all classes and IoU thresholds, thereby providing a comprehensive measure of detection accuracy (Jung *et al*., 2025). For example, mAP@0.5 evaluates detection performance at a fixed IoU threshold of 0.5, whereas mAP@0.5:0.95 averages results across multiple IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05, thereby offering a stricter and more robust evaluation of model performance. Together, these four metrics provide a rigorous framework for quantitatively assessing the effectiveness of YOLOv11 in detecting defective coffee beans. The mAP can be expressed as Equation (4), where N denotes the total number of object classes included in the evaluation, i represents the class index in the summation term, and AP_i refers to the average precision of the *i*-th class.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

3. RESULTS AND DISCUSSION

3.1. Model Training Performance

Figure illustrates the training and validation loss trajectories of the five YOLOv11 variants (n, s, m, l, and x) across 100 epochs. In general, all models exhibit a downward trend in loss values, indicating successful convergence without signs of divergence. The validation curves closely follow the training curves, suggesting that the models generalized well to unseen data without severe overfitting.

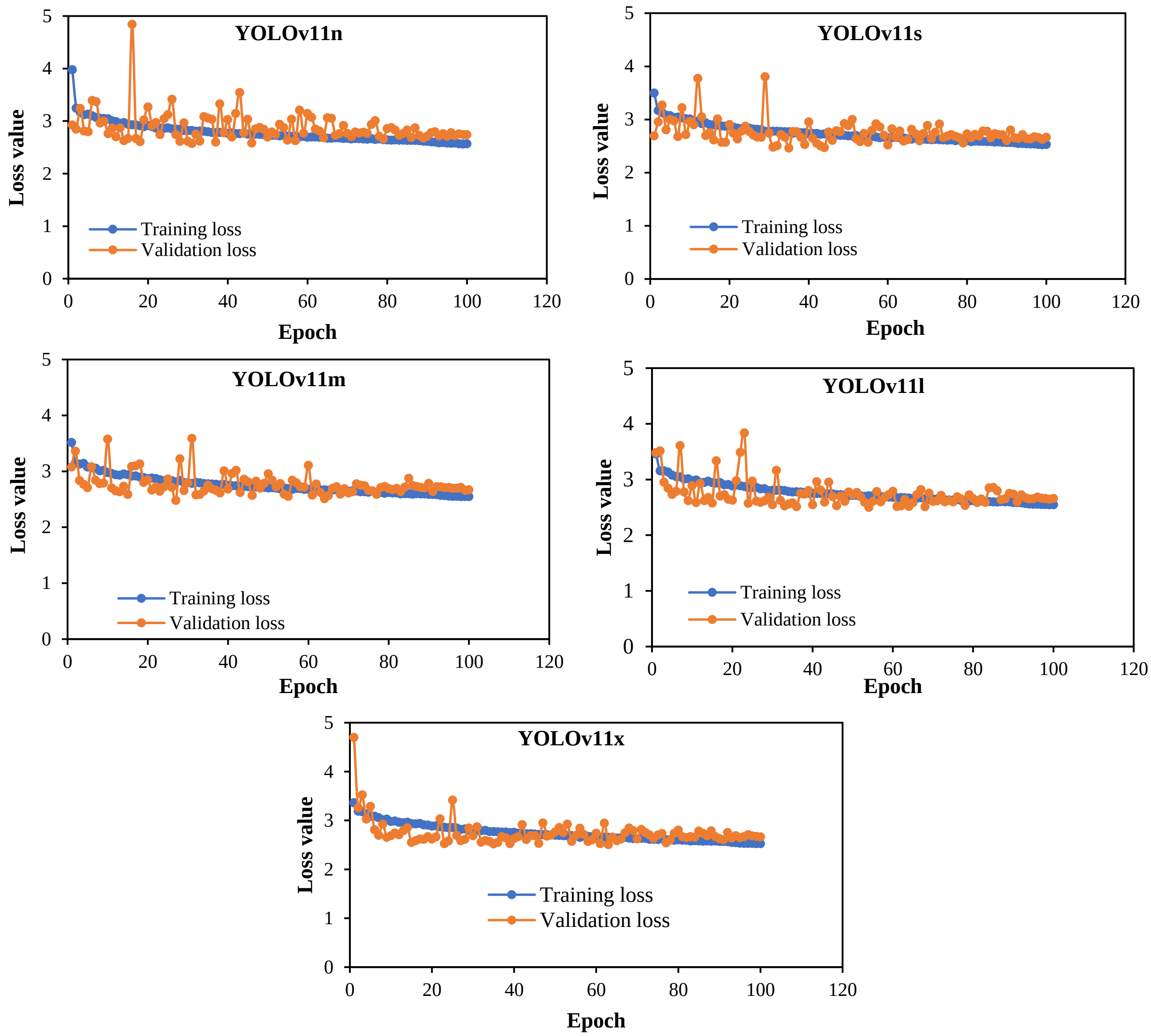


Figure 6. Training and validation loss curves for YOLOv11 variants (n, s, m, l, and x)

The lightweight YOLOv11n model shows a steady decline in both training and validation losses, stabilizing after approximately 60 epochs. Although several fluctuations are visible in the validation curve, the overall convergence pattern indicates effective learning. The YOLOv11s model exhibits a similar trajectory, with both losses gradually decreasing and stabilizing after around 70 epochs. Compared with YOLOv11n, YOLOv11s reaches slightly lower terminal loss values, reflecting improved representational capacity while maintaining computational efficiency.

The medium-sized YOLOv11m model shows relatively smooth convergence, with training and validation losses decreasing consistently. The validation curve closely tracks the training curve, indicating strong generalization capability and a reduced risk of overfitting. Likewise, YOLOv11l shows a progressive reduction in loss, with the training and validation curves remaining closely aligned throughout the process. Occasional fluctuations appear in the validation curve, but do not substantially affect the overall convergence pattern, suggesting robustness to data variability.

The largest model, YOLOv11x, achieves the lowest loss values among all variants. Both training and validation losses decline smoothly and converge to relatively low levels, highlighting its strong learning capacity. Nevertheless, minor oscillations in the validation curve suggest some sensitivity to dataset variation, which could potentially be mitigated through additional regularization.

Overall, the results indicate a positive relationship between model size and convergence performance. Larger models (YOLOv11l and YOLOv11x) attain lower loss values and show stable convergence, whereas smaller variants (YOLOv11n and YOLOv11s) offer a balance between computational efficiency and acceptable convergence behavior. These findings highlight the inherent trade-off between detection performance and computational demand, which is critical when selecting the most suitable model for real-time coffee bean defect detection.

Figure illustrates the training performance of the evaluated YOLOv11 variants based on precision, recall, mAP@0.5, and mAP@0.5:0.95 over 100 training epochs. As shown in panels (a)–(e), all models exhibit rapid convergence during the early phase of training, with most performance metrics stabilizing within the first 10–15 epochs, followed by minor fluctuations and a clear plateau toward the end of training. Across all variants, the precision and recall curves reach consistently high values and remain stable after convergence, indicating reliable detection behavior throughout the training process. Similarly, the mAP@0.5 curves indicate consistently high performance across all model scales, with only marginal differences between smaller and larger variants. In contrast, mAP@0.5:0.95 shows lower absolute values and greater variability, particularly for the lighter models, whereas the medium and larger variants (YOLOv11m, YOLOv11l, and YOLOv11x) exhibit improved stability and slightly higher mAP@0.5:0.95 values. Overall, the training curves confirm stable optimization and broadly comparable convergence behavior across all YOLOv11 variants, with incremental performance gains observed as model capacity increases from the nano to the extra-large configuration.

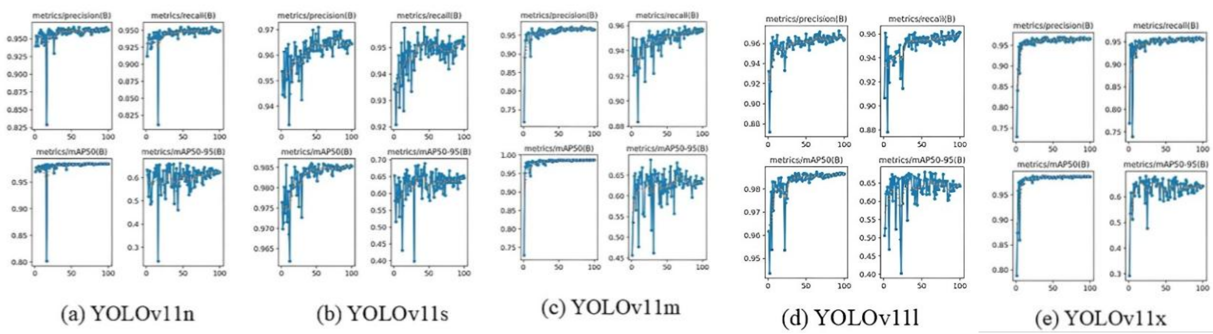


Figure 7. Training performance of YOLOv11 variants (n, s, m, l, and x)

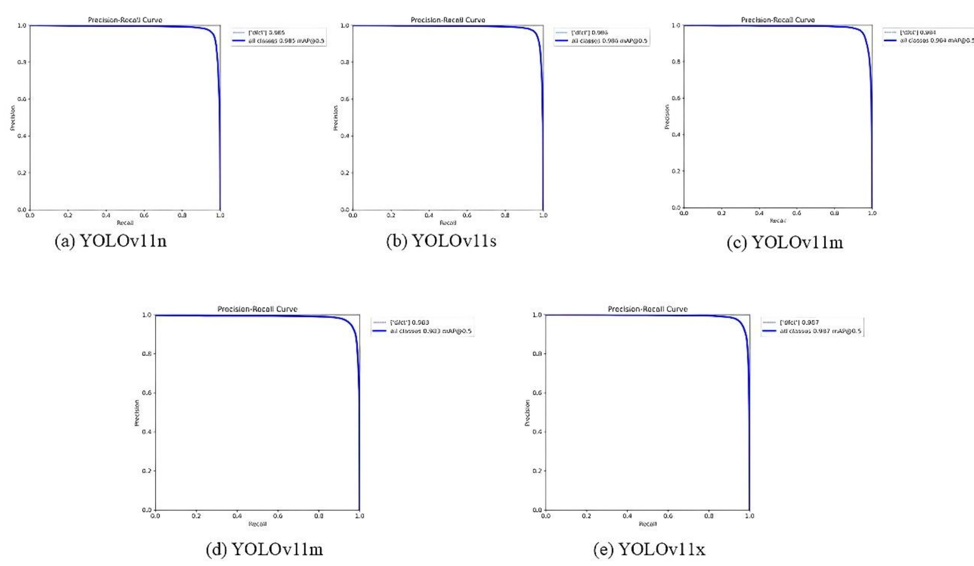
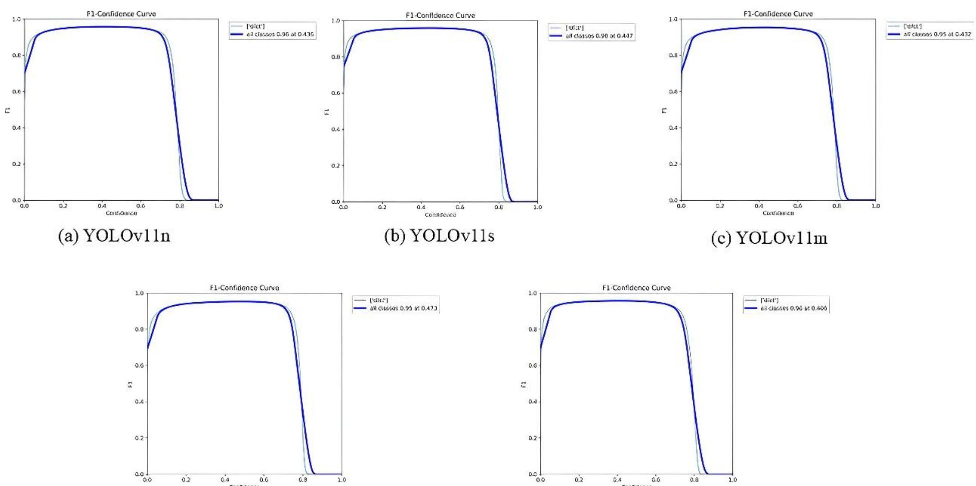


Figure 8. Precision-Recall curve of YOLOv11 variants (n, s, m, l, and x)

Figure presents the precision–recall (PR) curves for the five YOLOv11 variants. Across all panels, the curves remain close to the upper boundary (precision ≈ 1.0) over most of the recall range and decline sharply only near the highest recall region (0.95–1.0), indicating that high precision is maintained even as recall increases. The all-class summary in each panel shows high mAP@0.5 values for all variants, with YOLOv11s achieving the highest value (0.988), followed by YOLOv11x (0.987), YOLOv11n (0.985), and YOLOv11l (0.983), while YOLOv11m records the lowest value among the five (0.964). Overall, the PR curves exhibit closely aligned profiles with a consistently high area under the curve, indicating similar detection behavior across model scales under the evaluated conditions.

Figure shows the F1–confidence curves for the five YOLOv11 variants. Across all panels, the F1-score increases rapidly at low confidence thresholds, reaches a broad plateau, and then declines sharply as the confidence threshold approaches the upper range (0.8–1.0). The reported peak all-class F1-scores and their corresponding confidence thresholds are 0.96 at 0.415 for YOLOv11n, 0.98 at 0.447 for YOLOv11s, 0.97 at 0.432 for YOLOv11m, 0.96 at 0.473 for YOLOv11l, and 0.96 at 0.466 for YOLOv11x. Overall, the curves exhibit closely aligned shapes, with peak F1-scores clustered within 0.96–0.98 and optimal operating points occurring at mid-range confidence thresholds (0.41–0.47). Among the evaluated variants, YOLOv11s achieves the highest peak F1-score.



(a) YOLOv11n (b) YOLOv11s (c) YOLOv11m (d) YOLOv11l (e) YOLOv11x

Figure 9. F1-score curve of YOLOv11 variants (n, s, m, l, and x)

3.2. Evaluation Metrics Results

Table 3 summarizes the performance of the five YOLOv11 variants (n, s, m, l, and x) for defective coffee bean detection. Overall, all models achieved high precision (0.9540–0.9650) and recall (0.9440–0.9540), indicating reliable detection capability across different model scales. For the accuracy-based metrics, mAP@0.5 ranged from 0.9830 to 0.9870, whereas mAP@0.5:0.95 was lower, ranging from 0.6620 to 0.6890, reflecting the increased difficulty of accurate detection under stricter IoU thresholds.

Table 3. Performance comparison of YOLOv11 variants based on Precision, Recall, mAP, and F1-score for detection of defective coffee bean

Model	Precision	Recall	mAP50	mAP50-95	F1-score	Size (MB)	Training time (hours)	Inference time (ms)
YOLOv11n	0.9650	0.9520	0.9850	0.6620	0.9585	5.5	0.625	1.5
YOLOv11s	0.9640	0.9540	0.9860	0.6890	0.9590	19.2	1.036	3.3
YOLOv11m	0.9650	0.9440	0.9840	0.6880	0.9544	40.5	2.654	6.4
YOLOv11l	0.9540	0.9520	0.9830	0.6800	0.9530	51.2	3.055	8.3
YOLOv11x	0.9650	0.9520	0.9870	0.6881	0.9585	114.4	22.069	15.7

In terms of precision, YOLOv11n, YOLOv11m, and YOLOv11x each recorded the highest value of 0.9650, followed closely by YOLOv11s (0.9640), whereas YOLOv11l showed the lowest precision (0.9540). Recall was highest for YOLOv11s (0.9540), whereas YOLOv11n, YOLOv11l, and YOLOv11x each achieved 0.9520, and YOLOv11m performed slightly lower at 0.9440. For mAP@0.5, YOLOv11x achieved the highest score (0.9870), followed by YOLOv11s (0.9860), YOLOv11n (0.9850), YOLOv11m (0.9840), and YOLOv11l (0.9830). For mAP@0.5:0.95, the best result was obtained by YOLOv11s (0.6890), with YOLOv11x (0.6881) and YOLOv11m (0.6880) showing nearly identical performance, while YOLOv11l (0.6800) and YOLOv11n (0.6620) scored slightly lower.

The F1-score, which reflects the balance between precision and recall, was also highest for YOLOv11s (0.9590), followed by YOLOv11n and YOLOv11x (0.9585), YOLOv11m (0.9544), and YOLOv11l (0.9530). These results indicate that, although all variants performed consistently well, YOLOv11s and YOLOv11x delivered the strongest overall detection performance. Meanwhile, despite being the smallest model, YOLOv11n maintained competitive results, with only a slight reduction in the stricter mAP@0.5:0.95 metric.

3.3. Comparative Analysis of YOLOv11 Variants

A comparative analysis was conducted to evaluate the trade-off between detection accuracy and computational efficiency among the YOLOv11 variants. The results show that larger models generally achieved higher accuracy, whereas smaller variants offered faster inference and lower memory requirements. This trade-off is critical when selecting the most suitable model for real-time coffee bean defect inspection.

From a performance perspective, YOLOv11s achieved the best overall balance, recording the highest recall (0.9540), the best mAP@0.5:0.95 (0.6890), and the highest F1-score (0.9590). These results indicate that YOLOv11s not only detects defects with high precision but also maintains strong generalization under stricter IoU thresholds. This favorable balance may be attributed to its intermediate model scale, which appears to provide sufficient representational capacity for robust defect detection while maintaining lower computational demands than the larger variants. In other words, YOLOv11s offers a practical compromise between feature-learning capability and processing efficiency, allowing it to outperform the smallest variant under stricter localization criteria while avoiding the substantial model size, inference time, and training cost associated with YOLOv11x. YOLOv11x achieved the highest mAP@0.5 (0.9870) and matched the precision of YOLOv11n and YOLOv11m (0.9650). However, its computational demands, including a model size of 114.4 MB, an inference time of 15.7 ms, and a training duration exceeding 22 hours, make it less practical for real-time or resource-constrained environments.

In contrast, YOLOv11n demonstrated exceptional efficiency, with the smallest model size (5.5 MB) and the fastest inference speed (1.5 ms), while still achieving competitive detection performance (mAP@0.5 = 0.9850 and F1-score = 0.9585). This makes YOLOv11n an attractive option for deployment on edge devices with limited computational capacity, although its lower mAP@0.5:0.95 (0.6620) suggests reduced robustness under stricter IoU thresholds. The medium-sized variants, YOLOv11m and YOLOv11l, performed consistently well but did not surpass YOLOv11s or YOLOv11x in key detection metrics, while requiring more resources than YOLOv11s.

Overall, when both detection performance and computational efficiency are considered, YOLOv11s emerges as the most suitable model for real-time coffee bean defect inspection. This variant achieved the highest recall, the best mAP@0.5:0.95, and the strongest F1-score among all tested models, while maintaining a relatively small model size (19.2 MB) and low inference latency (3.3 ms). These characteristics indicate that YOLOv11s can deliver reliable and robust defect detection without imposing excessive computational burden, making it the most balanced choice for practical deployment in real-time inspection systems.

3.4. Performance Evaluation of YOLOv11s on Test Dataset

To further validate the reliability of YOLOv11s, the model was evaluated on a blind testing dataset comprising 500 coffee bean images. The evaluation included standard metrics such as precision (P), recall (R), mean Average Precision at an IoU threshold of 0.5 (mAP@0.5), mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP@0.5:0.95), F1-score, and inference time per image. The results are summarized in Table 4.

Table 4. Performance evaluation of YOLOv11s on testing dataset

P	R	mAP50	mAP50-95	F1-score	Inference
0.913	0.927	0.928	0.515	0.920	5.5

YOLOv11s achieved a precision of 0.913 and a recall of 0.927, indicating that the model was able to detect most defective coffee beans while maintaining a relatively low false-positive rate. The model recorded an mAP@0.5 of 0.928, demonstrating reliable detection performance under a lenient IoU threshold. However, the stricter evaluation based on mAP@0.5:0.95 yielded a lower value of 0.515, reflecting reduced localization accuracy under higher overlap requirements. The F1-score of 0.920 further confirms that YOLOv11s maintains a strong balance between precision and recall. The inference time was measured at 5.5 ms per image. Considering that the image acquisition system in this study operated at up to 30 fps and a conveyor speed of approximately 11 cm/s, this result supports the suitability of YOLOv11s for near real-time inspection under the present experimental conditions.

Notably, the blind-test results were lower than the training and validation results reported in Table 3, where YOLOv11s achieved an mAP@0.5:0.95 of 0.6890. This performance gap is commonly observed in deep learning and highlights the challenge of model generalization when analyzed on entirely unseen data. The reduction in mAP@0.5:0.95

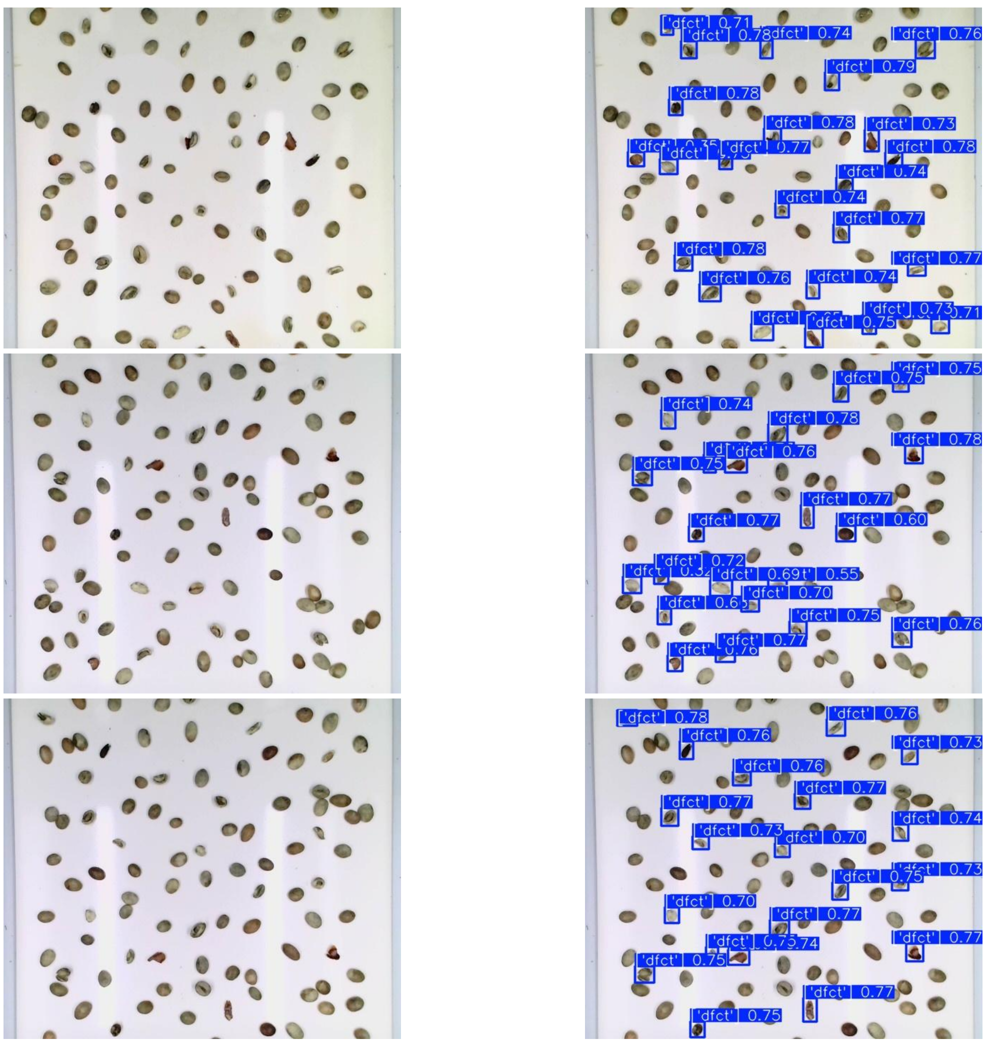


Figure 5. Representative inference results of YOLOv11s on testing dataset. Left: original input images. Right: detection outputs with bounding boxes and confidence scores

suggests that some defects in the blind dataset, particularly those influenced by lighting variation, bean orientation, or subtle surface irregularities, may not have been fully represented in the training data. Nevertheless, the consistently high recall and F1-score indicate that YOLOv11s retained strong defect detection capability even under stricter evaluation conditions.

In addition to the quantitative evaluation, a qualitative assessment of the inference results was performed. Representative examples are shown in Figure 5, where the left panels display the original test images and the right panels present the YOLOv11s predictions with bounding boxes and confidence scores. The model successfully identified defective beans, with confidence scores typically ranging from 0.60 to 0.79. Most defects were localized accurately, and non-defective beans were largely ignored, confirming the model's strong discriminative ability. A small number of false detections were observed, particularly for beans with borderline surface characteristics, which is consistent with the quantitative precision results.

Beyond overall detection performance, the ability of YOLOv11s to detect small defects was also examined. Detecting small objects is an important aspect of defect inspection because it directly reflects model robustness under challenging conditions. Based on the testing dataset, YOLOv11s was able to successfully localize defective beans of minimal size, with the smallest detected defect measuring approximately 5 mm and confidence scores above the detection threshold. Figure 5 presents representative examples of such small-object detections, in which YOLOv11s demonstrated reliable localization despite limited visual cues. These results indicate that the model retains strong discriminative capability even for small defects, further supporting its suitability for practical deployment in coffee bean quality control systems.

Overall, both quantitative and qualitative evaluations confirm that YOLOv11s remains a reliable and efficient model for coffee bean defect detection. Despite a reduction in mAP@0.5:0.95 relative to validation performance, the model demonstrated robustness in detecting defects in unseen data, maintained detection capability for small defects, and produced visual outputs that support its applicability in automated inspection pipelines.

3.5. Challenges, Limitations, and Future Research Opportunities toward Real-Time Inspection

Although all five YOLOv11 variants demonstrated excellent performance in detecting defective coffee beans, several limitations should be considered before extending these findings to fully operational real-time inspection systems. One of the main challenges identified in this study relates to model generalization under unseen conditions. While near-perfect mAP@0.5 values were achieved during training and validation, a noticeable performance decline particularly in mAP@0.5:0.95 was observed during blind testing. This indicates that localization accuracy is more sensitive to variations such as illumination changes, bean orientation, surface texture, and subtle defect morphology that were not fully represented in the training dataset. These findings highlight the need for more diverse and representative datasets to improve robustness under practical inspection conditions, although the overall detection capability remained strong.

Another important limitation lies in the simplified defect representation adopted in this study. All defective beans were grouped into a single class to enable a focused comparison of detection performance across the five YOLOv11 variants. Although this approach is appropriate for benchmarking model efficiency and accuracy, it does not capture the complexity of industrial coffee grading, where multiple defect categories with varying severity levels must be distinguished. In addition, the evaluation was conducted using a high-performance GPU environment, which may not fully reflect real deployment conditions on embedded or edge-based systems. Although lightweight models such as YOLOv11n and YOLOv11s showed low inference latency and compact model sizes, further investigation is needed to assess their stability, power efficiency, and long-term reliability on resource-constrained hardware. Practical challenges such as object overlap, occlusion, and environmental variability which were minimized in this study through controlled sample arrangement and illumination also remain important factors that may affect detection accuracy in real production lines.

Despite these limitations, the findings clearly demonstrate strong potential of YOLOv11-based models for real-time coffee bean defect inspection. All five variants achieved consistently high detection accuracy, confirming the effectiveness of the YOLOv11 architecture for non-destructive quality assessment. In particular, YOLOv11s emerged as the most balanced model, offering the best trade-off between detection accuracy, inference speed, and model size, which aligns directly with the conclusions of this study. Future research should build on these results by improving dataset diversity, exploring multi-class defect detection, evaluating deployment on edge platforms, and enhancing robustness to occlusion and environmental fluctuations. Within the defined scope of this work, the results provide a solid foundation for future progress toward scalable, reliable, and industry-ready real-time coffee bean inspection systems.

4. CONCLUSIONS

This study systematically evaluated the performance of five YOLOv11 variants (n, s, m, l, and x) for the non-destructive detection of defective coffee beans using a machine vision-based inspection system. All evaluated variants achieved consistently high detection accuracy, with mAP@0.5 values exceeding 0.98, confirming the effectiveness of the YOLOv11 architecture for visual defect detection in postharvest coffee quality assessment. YOLOv11s exhibited the most favorable balance between detection performance and computational efficiency, achieving the highest recall, mAP@0.5:0.95, and F1-score while maintaining low inference latency and a compact model size.

Beyond these performance outcomes, the scientific contribution of this study lies in three main aspects. First, it establishes a machine vision-based framework for non-destructive coffee bean defect inspection under controlled imaging conditions. Second, it provides a dedicated image dataset for coffee bean defect detection that supports systematic model training, validation, and blind-test evaluation. Third, it offers a comparative and application-oriented assessment of five YOLOv11 variants, enabling the identification of the most suitable model for practical real-time inspection within the defined experimental scope. In this context, the results identify YOLOv11s as the most suitable candidate for balancing accuracy, speed, and model compactness.

Although blind-test results revealed remaining challenges related to generalization, localization precision, and sensitivity to imaging variation, the findings provide a solid technical foundation for future research aimed at improving robustness, expanding defect granularity, and advancing toward scalable, real-time inspection systems for automated coffee quality control.

ACKNOWLEDGEMENT

The authors gratefully acknowledge financial support from the Institute for Research and Community Service (LPPM), University of Mataram, through the Penelitian Peningkatan Kapasitas research scheme funded by DIPA BLU, Fiscal Year 2025, under Contract Number 3095/UN18.L1.PP/2025.

AUTHOR CONTRIBUTION STATEMENT

Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
HK	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓
ISWS		✓				✓		✓		✓	✓	✓	✓	
HA		✓		✓		✓		✓		✓	✓	✓	✓	
SAM										✓		✓		
Suk										✓		✓		
Ans										✓		✓		
RS										✓		✓		
Mur										✓		✓		

C: Conceptualization	Fo: Formal Analysis	O: Writing - Original Draft	Fu: Funding Acquisition
M: Methodology	I: Investigation	E: Writing - Review & Editing	P: Project Administration
So: Software	D: Data Curation	Vi: Visualization	
Va: Validation	R: Resources	Su: Supervision	

REFERENCES

Adiwijaya, N.O., Sarno, R., & Wijaya, D.R. (2024). Coffee defects detection based on green bean images using YOLO architecture. In *Proceedings of the 2024 IEEE 2nd International Conference on Electrical Engineering, Computer and Information Technology (ICEECIT)* (pp. 314–319). https://doi.org/10.1109/ICEECIT63698.2024.10859988

Al Husaini, M., Raharja, A.R., Putra, V.H.C., & Lukmana, H.H. (2025). Enhanced plant disease detection using computer vision

YOLOv11: Pre-trained neural network model application. *Journal of Computer Networks, Architecture and High Performance Computing*, ***7***(1), 82–95. https://doi.org/10.47709/cnahpc.v7i1.5146

Al-Ghamdi, S., Alfaifi, B., Elamin, W., & Lateef, M.A. (2024). Advancements in coffee manufacturing: From dehydration techniques to quality control. *Food Engineering Reviews*, ***16***(4), 513–539. https://doi.org/10.1007/s12393-024-09383-5

Ali, M.M., Hashim, N., Aziz, S.A., & Lasekan, O. (2021). Quality inspection of food and agricultural products using artificial intelligence. *Advances in Agricultural and Food Research Journal*, ***2***(2). https://doi.org/10.36877/aafrj.a0000237

Alif, M.A.R., & Hussain, M. (2024). YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain. *ArXiv*, 1-31. http://arxiv.org/abs/2406.10139

Al-Mahish, M., & Alfayadh, R. (2024). The impact of coffee quality attributes and ratings on specialty coffee bean prices. *Beverage Plant Research*, ***4***(1), 0–0. https://doi.org/10.48130/bpr-0024-0029

Amadea, V., Rachmawati, E., Ferdian, E., & Akbar, M.N.S. (2024). Defect detection in Arabica green coffee beans based on grade quality. In *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence* (pp. 103–110). https://doi.org/10.1145/3669754.3669770

Anastácio, L.M., da Silva, M.deC.S., Debona, D.G., Veloso, T.G.R., Entringer, T.L., Bullergahn, V.B., da Luz, J.M.R., Moreli, A.P., Megumi, M.C., & Pereira, L.L. (2023). Relationship between physical changes in the coffee bean due to roasting profiles and the sensory attributes of the coffee beverage. *European Food Research and Technology*, ***249***(2), 327–339. https://doi.org/10.1007/s00217-022-04118-4

Ardian, Y., Irawan, N.D., Sutoko, S., Astawa, I.N.G.A., Purnama, I.B.I., & Dwiyanto, F.A. (2024). A novel approach to defect detection in Arabica coffee beans using deep learning: Investigating data augmentation and model optimization. *Knowledge Engineering and Data Science*, ***7***(1), 117–127. https://doi.org/10.17977/um018v7i12024p117-127

Arwatchananukul, S., Xu, D., Charoenkwan, P., Moon, S.A., & Saengrayap, R. (2024). Implementing a deep learning model for defect classification in Thai Arabica green coffee beans. *Smart Agricultural Technology*, ***9***, 100680. https://doi.org/10.1016/j.atech.2024.100680

Badgujar, C.M., Poulose, A., & Gan, H. (2024). Agricultural object detection with You Only Look Once (YOLO) algorithm: A bibliometric and systematic literature review. *Computers and Electronics in Agriculture*, ***223***, 109090. https://doi.org/10.1016/j.compag.2024.109090

Bumbálek, R., Ufitikirezi, J.deD.M., Umurungi, S.N., Zoubek, T., Kuneš, R., Stehlík, R., & Bartoš, P. (2025). Computer vision in precision livestock farming: Benchmarking YOLOv9, YOLOv10, YOLOv11, and YOLOv12 for individual cattle identification. *Smart Agricultural Technology*, ***12***, 101208. https://doi.org/10.1016/j.atech.2025.101208

Chandu, B., Surendran, R., & Selvanarayanan, R. (2025). To instant of coffee beans using K-nearest algorithm over clustering for quality and sorting process. In *Proceedings of the 2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD)* (pp. 1–7). https://doi.org/10.1109/ITIKD63574.2025.11005039

Chang, S.J., & Liu, K.H. (2024). Multiscale defect extraction neural network for green coffee bean defects detection. *IEEE Access*, ***12***, 15856–15866. https://doi.org/10.1109/ACCESS.2024.3356596

Chen, F., Zhang, Y., Fu, L., Hua, R., Zhang, Q., & Bi, S. (2025). A comparative review of the next-generation YOLO models: YOLOv10 and YOLO11. *Journal of Computer Science and Artificial Intelligence*, ***3***(2), 1–6. https://doi.org/10.54097/22zsmc10

Chen, S.-Y., Chiu, M.-F., & Zou, X.-W. (2022). Real-time defect inspection of green coffee beans using NIR snapshot hyperspectral imaging. *Computers and Electronics in Agriculture*, ***197***, 106970. https://doi.org/10.1016/j.compag.2022.106970

García, M., Candelo-Becerra, J.E., & Hoyos, F.E. (2019). Quality and defect inspection of green coffee beans using a computer vision system. *Applied Sciences*, ***9***(19), 4195. https://doi.org/10.3390/app9194195

Gope, H.L., Fukai, H., Ruhad, F.M., & Barman, S. (2024). Comparative analysis of YOLO models for green coffee bean detection and defect classification. *Scientific Reports*, ***14***(1), 28946. https://doi.org/10.1038/s41598-024-78598-7

Gosalvitr, P., Cuéllar-Franca, R.M., Smith, R., & Azapagic, A. (2024). Environmental and economic sustainability assessment of the production and consumption of different types of coffee in the UK. *Sustainable Production and Consumption*, ***49***, 144–162. https://doi.org/10.1016/j.spc.2024.06.004

He, L., Zhou, Y., Liu, L., Cao, W., & Ma, J. (2025). Research on object detection and recognition in remote sensing images based on YOLOv11. *Scientific Reports*, ***15***(1), Article 14032. https://doi.org/10.1038/s41598-025-96314-x

Hemamalini, V., Rajarajeswari, S., Nachiyappan, S., Sambath, M., Devi, T., Singh, B.K., & Raghuvanshi, A. (2022). Food quality inspection and grading using efficient image segmentation and machine learning-based system. *Journal of Food Quality*,

2022, 5262294. https://doi.org/10.1155/2022/5262294

Hidayatullah, P., Syakrani, N., Sholahuddin, M.R., Gelar, T., & Tubagus, R. (2025). YOLOv8 to YOLO11: A comprehensive architecture in-depth comparative review. *arXiv*, 1-22. http://arxiv.org/abs/2501.13400

Ji, Y., Xu, J., & Yan, B. (2024). Coffee green bean defect detection method based on an improved YOLOv8 model. *Journal of Food Processing and Preservation*, ***2024***(1), 2864052. https://doi.org/10.1155/2024/2864052

Jung, E., & Nam, D. (2025). Lightweight YOLO-based real-time fall detection using feature map-level knowledge distillation. *ICT Express*, ***11***(6), 1152-1161. https://doi.org/10.1016/j.icte.2025.08.009

Katsamenis, I., Karolou, E.E., Davradou, A., Protopapadakis, E., Doulamis, A., Doulamis, N., & Kalogeras, D. (2023). TraCon: A novel dataset for real-time traffic cones detection using deep learning. In *Proceedings of the conference* (pp. 382–391). https://doi.org/10.1007/978-3-031-17601-2_37

Khanam, R., & Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. *arXiv*, 1-9. http://arxiv.org/abs/2410.17725

Kishor, R. (2024). Performance benchmarking of YOLOv11 variants for real-time delivery vehicle detection: A study on accuracy, speed, and computational trade-offs. *Asian Journal of Research in Computer Science*, ***17***(12), 108–122. https://doi.org/10.9734/ajrcos/2024/v17i12532

Li, Z., Wang, L., You, D., & Chen, L. (2025). LDV-YOLO: A lightweight improved YOLOv11 for photovoltaic cell defect detection. *Contemporary Mathematics*, ***6***(5), 6259–6274. https://doi.org/10.37256/cm.6520257513

Liang, C.-S., Xu, Z.-Y., Zhou, J.-Y., Yang, C.-M., & Chen, J.-Y. (2023). Automated detection of coffee bean defects using multi-deep learning models. In *Proceedings of the 2023 VTS Asia Pacific Wireless Communications Symposium (APWCS)* (pp. 1–5). https://doi.org/10.1109/APWCS60142.2023.10234059

Lin, Y., Ma, J., Wang, Q., & Sun, D.-W. (2023). Applications of machine learning techniques for enhancing nondestructive food quality and safety detection. *Critical Reviews in Food Science and Nutrition*, ***63***(12), 1649–1669. https://doi.org/10.1080/10408398.2022.2131725

Luis, V.A.M., Quinones, M.V.T., & Yumang, A.N. (2022). Classification of defects in Robusta green coffee beans using YOLO. In *Proceedings of the 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)* (pp. 1–6). https://doi.org/10.1109/IICAIET55139.2022.9936831

Matias, G.C., Borém, F.M., Alves, A.P.C., Haeberlin, L., Santos, C.M.dos, & Andrade, E.T.de. (2025). Impact of initial sensory quality of specialty natural coffee during storage. *Anais da Academia Brasileira de Ciências*, ***97***(2). https://doi.org/10.1590/0001-3765202520241042

Murat, A.A., & Kiran, M.S. (2025). A comprehensive review on YOLO versions for object detection. *Engineering Science and Technology, an International Journal*, ***70***, 102161. https://doi.org/10.1016/j.jestch.2025.102161

Oncu, E. (2025). Withdrawn: Deep learning framework for coffee quality assessment via YOLOv8n object detection of bean defects. *Preprints*. https://doi.org/10.20944/preprints202509.0643.v1

Rasheed, A.F., & Zarkoosh, M. (2025). YOLOv11 optimization for efficient resource utilization. *ArXiv*, 1-14. http://arxiv.org/abs/2412.14790

Samudra, M.N., & Rachmawati, E. (2025). Deep learning-based defect detection in Arabica green coffee beans using LSKNet. In *Proceedings of the 2025 International Conference on Data Science and Its Applications (ICoDSA)* (pp. 692–697). https://doi.org/10.1109/ICoDSA67155.2025.11157423

Shen, C., Wang, R., Nawazish, H., Wang, B., Cai, K., & Xu, B. (2024). Machine vision combined with deep learning–based approaches for food authentication: An integrative review and new insights. *Comprehensive Reviews in Food Science and Food Safety*, ***23***(6), e70054. https://doi.org/10.1111/1541-4337.70054

Thai, H.D., Ko, H.J., & Huh, J.H. (2024). Coffee bean defects automatic classification realtime application adopting deep learning. *IEEE Access*, ***12***, 126503–126517. https://doi.org/10.1109/ACCESS.2024.3452552

Wang, P., Tseng, H.-W., Chen, T.-C., & Hsia, C.-H. (2021). Deep convolutional neural network for coffee bean inspection. *Sensors and Materials*, ***33***(7), 2299–2310. https://doi.org/10.18494/SAM.2021.3277

Xia, Y., Nguyen, M., & Yan, W.Q. (2023). A real-time kiwifruit detection based on improved YOLOv7. In *Lecture Notes in Computer Science* (Vol. 13836, pp. 48–61). https://doi.org/10.1007/978-3-031-25825-1_4

Zhao, Q., & Zhu, J. (2025). An improved YOLOv11 architecture with multi-scale attention and spatial fusion for fine-grained residual detection. *Results in Engineering*, ***27***, 107061. https://doi.org/10.1016/j.rineng.2025.107061