Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1383~1390
e-ISSN: 2548-6861

Comparative Analysis of Deep Learning Architectures for Coffee Tree Detection from Aerial Imagery

Alya Khairunnisa Rizkita [1]*, Andre Febrianto [2]*, Amirul Iqbal [3]*, Miranti Verdiana [4]*, Muhammad Habib Algifari [5]*, Eko Dwi Nugroho [6]*
* Teknik Informatika, Institut Teknologi Sumatera
alya.rizkita@if.itera.ac.id [1], andre.febrianto@if.itera.ac.id [2], amirul.iqbal@if.itera.ac.id [3], miranti.verdiana@if.itera.ac.id [4], muhammad.algifari@if.itera.ac.id [5], eko.nugroho@if.itera.ac.id [6]

Article Info

Article history:

Received 2026-01-27
Revised 2026-03-15
Accepted 2026-04-10

Keyword:

Deep Learning,
YOLO,
Aerial Images,
Augmentation.

ABSTRACT

Coffee cultivation plays a vital economic role globally, supporting millions of livelihoods. Traditional manual enumeration methods for crop monitoring are time-intensive, costly, and prone to errors, particularly on large-scale farms. This study addresses the need for automated coffee tree detection systems by systematically evaluating five state-of-the-art deep learning architectures: YOLOv8 (nano, small, medium), Faster R-CNN, and EfficientDet. Using a dataset of 1,500 high-resolution aerial images from coffee plantations in Lampung, we investigated four critical aspects: optimal object detection architecture, effective augmentation strategies, minimum data requirements, and error patterns. Results demonstrate that YOLOv8n achieves superior performance with 95.98% mAP@0.5, outperforming larger variants and two-stage detectors. Basic augmentation techniques proved most effective, with mAP@0.5 of 96.13%, surpassing aggressive strategies like mosaic and mixup that disrupted the spatial structure of the plantations. Data efficiency analysis revealed that 750 images (50% of the dataset) achieved 99.55% of peak performance, enabling cost-effective deployment in resource-constrained scenarios. Error analysis indicated that false positives were the primary challenge, which is addressable through confidence threshold calibration. These findings provide evidence-based guidelines for practitioners, demonstrating that compact architectures with moderate augmentation can achieve high accuracy with limited data, facilitating the practical deployment of precision agriculture technologies in coffee cultivation.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Coffee stands as one of the most valuable commodities crops globally, generating an annual industrial revenue exceeding $300 billion. Additionally, coffee sustains the livelihoods of millions across various nations [1]. Accurate data management for coffee cultivation is essential for effective crop oversight, enabling precise land records, optimal allocation of raw materials, and strategic long-term planning [2], [3]. Conventional methods typically involve manual crop enumeration by farmers, which is time-intensive and costly. Moreover, these results are often inconsistent and prone to human error, especially on extensive farms [4]. As the demand for coffee continues to escalate while arable land remains constrained, there is an increasing imperative to enhance cultivation practices through the adoption of precision agriculture technologies [5], [6].

Monitoring agricultural land through remote sensing has become increasingly accessible, owing to the availability of affordable drones and high-quality imaging technologies [7]. Coupled with the benefits of deep learning within computer vision, this technology facilitates plant detection and enumeration. Automated systems can process detections of thousands of plant species within minutes, ensuring consistent measurements and enabling data-driven decision-making [8], [9], [10]. Nonetheless, translating this technology into a

practical, deployable solution for coffee cultivation requires a systematic investigation of detection architectures, optimization strategies, and real-world constraints.

Despite the considerable potential of deep learning for object detection in agricultural fields, several significant challenges remain in detecting coffee plants from remote sensing imagery. Firstly, coffee trees in remote sensing images present numerous unique challenges, including their small size, dense arrangement, and variable appearances influenced by factors such as plant growth stages, lighting conditions, and climate or weather variations [11]. Secondly, although various object detection architectures have been developed in recent years from traditional two-stage detectors such as Faster R-CNN [12] to contemporary one-stage approaches such as YOLO [13] and EfficientDet [14], there is currently a lack of comprehensive comparison regarding their performance in the specific context of coffee plant detection. Each architecture offers distinct advantages relating to accuracy, inference speed, and model complexity, yet practitioners lack clear guidance on selecting the most suitable approach for agricultural deployment. Thirdly, data augmentation, a crucial step for enhancing model generalization and robustness [15], has not been systematically evaluated within remote sensing image processing, particularly for coffee plants. Techniques such as mosaic augmentation and mix-up, despite their effectiveness in general image processing, may not be directly applicable to remote sensing imagery. Fourth, the minimal data requirements for achieving acceptable detection performance remain uncertain, leaving practitioners unsure about the appropriate application of augmentation techniques. Lastly, understanding the causes of model failures in pattern recognition is an important component of systematic error analysis. Several studies indicate that most research on object detection within agriculture concentrates on evaluating a single architecture without comprehensive comparative analyses [16], [17]. While some research confirms the potential of deep learning for tree detection [11], [18], these studies typically employ only a single detection framework and do not explore alternative methods. Moreover, computer vision research on coffee plants has predominantly focused on disease detection [19, 20] and fruit ripeness classification [21], with limited emphasis on detecting and quantifying individual trees. Only a few studies provide detailed analyses of crop tree detection, addressing issues such as data efficiency, augmentation techniques, and error patterns, which are essential for practical applications [22].

The completeness of existing research is hampered by severely limited resources in many coffee-producing regions. In the absence of a well-defined framework for minimum data requirements, the optimal architectural model, and effective augmentation strategies, researchers continue to rely on trial-and-error methods, resulting in inefficient use of valuable resources. Moreover, the lack of comprehensive error analysis highlights implementation difficulties that arise during production, and mitigation efforts also incur high costs. Consequently, this study proposes several research avenues to be examined empirically, guided by the subsequent research questions:

1. Which object detection architecture offers the optimal balance for coffee tree identification using aerial imagery?
2. Which data augmentation strategy enhances detection performance for aerial coffee plantation images?
3. What is the minimum quantity of training data necessary to attain satisfactory detection results?
4. What are the primary error patterns and limitations of the model?

By systematically addressing these inquiries, we aim to provide evidence-based recommendations that facilitate the efficient and cost-effective deployment of coffee tree detection systems. Additionally, we seek to establish methodological foundations that are applicable to a broader range of agricultural monitoring tasks.

The primary contributions of this study are fivefold. First, we conduct the first systematic evaluation of five state-of-the-art object detection architectures for coffee tree detection. Secondly, we illustrate that fundamental augmentation strategies outperform more aggressive techniques such as mosaic and mix-up, thereby challenging prevailing practices in natural image augmentation. Thirdly, our data-efficiency analysis indicates that 750 images achieve 99.55% of peak performance, providing practical guidelines for resource-constrained scenarios. Fourthly, comprehensive error analysis achieves a 95.44% recall, facilitating effective threshold tuning for deployment. Finally, we synthesize these findings into a comprehensive deployment framework with actionable recommendations concerning model selection, augmentation, and configuration.

This paper is structured as follows. Section II reviews related work on object detection architectures, agricultural AI applications, and data augmentation techniques. Section III outlines our methodology, covering dataset preparation, model architectures, training settings, and evaluation metrics. Section IV details experimental results from four studies: architecture comparison, augmentation analysis, data efficiency, and error analysis. Section V discusses the implications of our findings, deployment considerations, and limitations. Finally, Section VI summarizes our contributions and suggests future research directions.

II. RELATED WORK

A. Object Detection

Object detection involves identifying and localizing objects in an image by predicting both the class label and the bounding box coordinates for each object [23]. In contrast to classification, which assigns a single label to each image, object detection can detect multiple objects simultaneously within a single image [24]. Object detection methodologies

have evolved from handcrafted feature-based techniques, such as HOG [25] and DPN [26], to deep learning approaches. Krizhevsky et al. [27] demonstrated the efficacy of convolutional neural networks (CNNs) in object detection, achieving markedly superior performance compared to traditional methods. This advancement laid the foundation for the ongoing development of deep learning-based object detection techniques.

Modern architectures fall into two paradigms: two-stage and one-stage detectors. Two-stage detectors (R-CNN family [12]) separate region proposal generation from classification, prioritizing accuracy. One-stage detectors (YOLO [13], SSD [28]) directly predict bounding boxes and classes in a single pass, achieving real-time performance with competitive accuracy. However, most research is conducted on general datasets with diverse, unstructured scenes.

B. R-CNN

The Region-based Convolutional Neural Network (R-CNN) family represents the pioneering approach to integrating deep learning into object detection, establishing the two-stage detection paradigm that separates region proposal generation from object classification and localization refinement. R-CNN, proposed by Girshick et al. [12], marked the first successful application of CNNs to object detection, achieving a 30% relative improvement over prior methods on the PASCAL VOC dataset. The architecture operates in three stages: generating approximately 2,000 region proposals per image using selective search; extracting fixed-length feature vectors from each proposal using a CNN; and classifying each region with class-specific linear SVMs, followed by bounding box regression. Despite its breakthrough performance, R-CNN suffered from severe computational inefficiency: processing each proposal independently through the CNN required days of GPU training, and inference took 47 seconds per image, rendering it impractical for real-world applications [12].

Fast R-CNN [29], introduced by Girshick in 2015, addressed R-CNN's inefficiency through architectural redesign. Rather than extracting features from each proposal independently, Fast R-CNN processes the entire image through a CNN once, then projects proposals onto the resulting feature map. A Region of Interest (RoI) pooling layer extracts fixed-size feature vectors from arbitrary-sized proposals, which are then classified and refined by fully connected layers. This single-pass design reduced training time from 84 hours to 9 hours and inference time from 47 seconds to 2.3 seconds per image [29]. However, Fast R-CNN still relied on selective search for proposals, which remained the computational bottleneck at 2 seconds per image.

C. YOLO

YOLO (You Only Look Once) revolutionized object detection by reformulating it as a single regression problem, directly predicting bounding boxes and class probabilities in a single network evaluation, thereby enabling real-time performance while maintaining competitive accuracy [13]. YOLOv1, introduced by Redmon et al. in 2016, represented a paradigm shift by dividing the input image into an S×S grid, achieving 45 FPS on Titan X GPU over 100 times faster than Faster R-CNN with 63.4% mAP on PASCAL VOC 2007, although it struggled with small object detection due to coarse spatial resolution [13]. YOLOv2 addressed these limitations through architectural improvements, including batch normalization, anchor boxes, and finer 13×13 feature maps, thereby achieving 78.6% mAP while maintaining frame rate [30]. YOLOv4 aggregated numerous architectural innovations, including CSPDarknet53, SPP, and PANet, along with advanced training strategies such as mosaic data augmentation, achieving 43.5% mAP@0.5 on MS COCO while maintaining 65 FPS on a V100 GPU [31]. YOLOv8, the latest iteration, features anchor-free detection heads and enhanced feature fusion via C2f modules, and the YOLOv8n variant achieves 37.3% mAP@0.5:0.95 with only 3.2M parameters at 80+ FPS on an A100 GPU, making it particularly suitable for edge deployment [32]. The YOLO family's evolution demonstrates continuous improvement in the speed-accuracy trade-off.

D. EfficientDet

EfficientDet, proposed by Tan et al. [14] in 2020, addresses the challenge of systematically scaling object detectors for different resource constraints by applying principled compound scaling across all network dimensions simultaneously. The key innovation is the Bidirectional Feature Pyramid Network (BiFPN), which enhances multi-scale feature fusion by incorporating weighted bidirectional connections and removing low-contribution nodes [14]. EfficientDet employs EfficientNet as its backbone and extends compound scaling principles to the entire detection network, jointly scaling backbone, BiFPN, and prediction network dimensions using a unified rule [14]. The architecture offers eight model variants (D0-D7) spanning different computational budgets. EfficientDet-D0, with 3.9M parameters, achieves 33.8% mAP@0.5:0.95 on MS COCO at 33 FPS, while D7 achieves 52.2% mAP at 5.5 FPS [14]. In resource-constrained scenarios, D0's 4M parameters and 98M FLOPs enable deployment on mobile devices and edge hardware, although its superiority for coffee tree detection remains to be empirically evaluated [14].

III. METHODOLOGY

A. Dataset

The dataset comprises 1,500 high-resolution aerial images of coffee plantations, captured using unmanned aerial vehicles (UAV) equipped with multispectral cameras. These images were collected from coffee plantations in Lampung, encompassing diverse plantation conditions, including varying tree ages, canopy densities, and lighting conditions. The image size is 1600×1300 pixels during daylight hours,

under clear to partially cloudy skies, ensuring high image quality while capturing natural illumination variations.

Figure 1. Dataset Sample

The dataset contains two object classes: "coffee," representing individual coffee trees, and "other," representing additional vegetation types occasionally present within plantations. Each coffee tree was annotated with a rectangular bounding box enclosing its canopy from an overhead perspective. Annotations were performed using the LabelImg annotation tool. Quality control procedures included random sampling and cross-validation of 10% of the annotations to verify consistency, achieving an inter-annotator agreement exceeding 95%. At the beginning of training, the dataset was randomly split into training (70%), validation (15%), and test (15%) sets, as shown in Table 1.

TABLE 1.
DATASET SPLIT

Split	Images	Object
Training (70%)	1050	35572
Validation (15%)	225	7623
Test (15%)	225	7622
Total (100%)	1500	50817

B. Model Architectures

We evaluate five sophisticated object detection architectures embodying various design paradigms: the YOLOv8 family (nano, small, and medium variants), Faster R-CNN with ResNet-50-FPN backbone, and EfficientDet. These models span a wide range of speed-accuracy-size trade-offs, enabling comprehensive analysis suitable for diverse deployment scenarios. The specifications of the model architectures are detailed in Table 2.

TABLE 2.
MODEL ARCHITECTURE SPECIFICATIONS

Model	Type	Parameters	Size (MB)
YOLOv8n	One stage	3.2 M	6.2
YOLOv8s	One stage	11.2 M	22.5
YOLOv8m	One stage	25.9 M	52
Faster R-CNN	Two stages	41.8 M	167
EfficientDet	One stage	3.9 M	15.6

C. Training Configuration

To ensure equitable comparison across different architectures, we utilized matched training protocols with consistent hyperparameters wherever feasible, while adhering to architecture-specific requirements. All models were trained for 100 epochs, with early stopping triggered after 50 epochs based on validation mAP. The training process was conducted on a single NVIDIA RTX 4060 GPU with 8GB of VRAM. We employed the AdamW optimizer with an initial learning rate of 0.001, a final learning rate of 0.00001, a weight decay of 0.0005, and a momentum of 0.937 for all models. The learning rate schedule used cosine annealing with a 3-epoch warmup period, starting from an initial learning rate of 0.0001. The loss function weights for the YOLO models were set to box=7.5, class=0.5, and distribution focal loss (DFL)=1.5, balancing localization and classification objectives. For Faster R-CNN and EfficientDet, the default loss configurations from their respective frameworks were employed.

Due to memory constraints and architectural differences, certain parameters were adjusted for each model.

1. Batch Size: YOLOv8 models employed a batch size of 16, optimizing GPU memory utilization and gradient estimation quality. Conversely, Faster R-CNN, which requires more memory due to Region of Interest (RoI) operations, used a batch size of 4. The EfficientDet model used a batch size of 8, balancing the two.
2. Input Resolution: Both YOLOv8 and Faster R-CNN processed images at 640×640 pixels, a resolution that maintains a balance between preserving detail and computational efficiency. The EfficientDet model used an input resolution of 512×512 pixels, in accordance with its original design and compound scaling methodology.

The experimental scenario comprises four categories of experiments: object detection architecture, augmentation implementation strategy, efficiency metrics, and error analysis.

D. Evaluation metrics

The evaluation metrics cover detection accuracy, efficiency, and error characterization. Mean Average Precision (mAP) serves as the primary metric for detection accuracy, following standard practices in object detection research [13]. mAP summarizes the precision-recall curve across all classes and images, providing a single score that balances detection completeness (recall) and correctness (precision). There are two mAP variants following standard object detection evaluation practices. mAP@0.5 uses an IoU threshold of 0.5 to determine correct detections. mAP@0.5:0.95 averages precision across IoU thresholds from 0.5 to 0.95 in steps of 0.05, providing stricter evaluation of localization quality as used in MS COCO benchmarks [13].

III. Result and Discussion

A. Experiment Results

This section discusses the quantitative results from the four experiments conducted: architecture comparison, analysis of augmentation techniques, data efficiency evaluation, and error distribution. A complete interpretation is presented in Table 3, which describes the results of the performance comparison of 5 object detection architectures. The results of this experiment show that YOLOv8n outperforms the mAP@0.5 of 95.98%, demonstrating that the YOLOv8n variant is the most effective architecture for coffee plant detection. Furthermore, there is another variant, YOLOv8s, which has an mAP@0.5 of 95.83%, a maximum F1-Score of 91.01%, and an mAP@0.5:0.95 of 78.85%. YOLOv8m gets the highest mAP@0.5:0.95 value of 79.20%. Faster R-CNN, compared to other architectures, does not achieve better performance, with a nAP@0.5 of 91.98%, which is almost 4% lower than YOLOv8n. EfficientDet, like Faster R-CNN, also achieved the lowest performance with a mAP@0.5 value of 88.00%. In the architecture performance comparison experiment, YOLO dominated overall, with YOLOv8n having the best performance.

Table 4 presents the experimental results of applying augmentation techniques to YOLOv8n. 4 augmentation techniques were used in this experiment. Basic technique with a fairly moderate transformation with HSV adjustment (hue ±0.015, saturation 0.7, value 0.4), rotation ±10°, translation 10%, scale ±50%, and horizontal flip 50%. Medium technique with higher parameter rotation ±15°, translation 20%, scale ±90%, and mosaic 50%. Heavy technique with parameters rotation ±20°, shear ±5°, mosaic 100%, mixup 15%. The basic technique achieves the best performance, with mAP@0.5 of 96.13% and mAP@0.5:0.95 of 77.16%, showing a greater improvement than the baseline that does not apply augmentation. The medium technique achieved an mAP@0.5 value of 95.60%, which is still below the basic augmentation. And the heavy technique, which is the most complete augmentation, achieved an mAP@0.5 of 94.87%, which is still far below that of the basic technique. Overall, all augmentation techniques improve the architecture's performance, as evidenced by the mAP@0.5, which is higher than the baseline. However, the superiority of basic augmentation across all metrics establishes moderate geometric and photometric transformations as optimal for aerial coffee imagery, while mosaic and mixup strategies prove counterproductive.

Table 5 presents the results of the data-efficiency experiment across four levels: 25% with 375 images, 50% with 750 images, 75% with 1125 images, and 100% with 1500 images. The performance with 25% data usage is 94.94%, 50% is 95.70%, 75% is 95.97%, and the overall is 96.13%. This pattern indicates that as data volume and diversity increase, the performance of the learning architecture also improves.

(a) (b)

Figure 2. (a) Baseline; (b) Detection Result

Table 6 categorizes detection outcomes for 225 test images containing 7,856 ground-truth objects. True Positives dominate with 7,395 instances (80.08%), indicating strong detection capability. False Positives represent the primary challenge, accounting for 1,379 instances (14.93%), approximately 4 times as frequent as False Negatives (353, 3.82%). This pattern reveals the model's aggressive detection strategy, favouring over-detection to avoid missing objects. Localization Errors and Misclassifications are minimal, confirming accurate bounding box placement and clear class distinction when objects are detected. High recall affirms the model's effectiveness in detecting coffee trees, while moderate precision suggests potential for improvement through threshold tuning. The predominant False Positive category indicates that addressing over-detection, rather than missed objects, should be the primary focus for deployment optimization. An example of an incorrect detection result can be seen in Figure 2(b) with the label in Figure 2(a)

Table 3.
Architecture Comparation

Model Size (MB)	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1-Score
YOLOv8n – 6.2	**95.98%**	77.76%	93.09%	88.26%	90.61%
YOLOv8s – 22.5	95.83%	78.85%	**93.83%**	88.36%	**91.01%**
YOLOv8m – 52	95.66%	**79.20%**	92.43%	**88.98%**	90.67%
Faster R-CNN – 167	91.98%	68.50%	91.98%	88.47%	90.19%
EfficientDet – 15.6	88.00%	65.00%	89.00%	85.00%	86.95%

TABLE 4.
AUGMENTATION STRATEGY

Augmentation Strategy	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1-Score
Baseline	90.15%	65.78%	90.16%	82.24%	86.01%
Basic	96.13%	77.16%	91.71%	89.74%	90.71%
Medium	95.60%	73.16%	91.10%	88.62%	89.84%
Heavy	94.87%	72.60%	90.87%	87.31%	89.05%

TABLE 5.
DATA EFFICIENCY

Data Size (%)	Images	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1-Score
25%	375	94.94%	74.35%	91.94%	87.07%	89.44%
50%	750	95.70%	76.60%	91.30%	88.35%	89.80%
75%	1125	95.97%	77.65%	92.29%	88.28%	90.24%
100%	1500	96.13%	77.15%	91.71%	89.74%	90.71%

TABLE 6.
ERROR DISTRIBUTION

Category	Count	Percentage
True Positives	7395	80.08%
False Positives	1379	14.93%
False Negatives	353	3.82%
Localization Errors	104	1.13%
Misclassifications	4	0.04%
Total	9235	100%

The final experiment regarding error distribution is shown in Table 6. Based on the experiment, 225 test images contain 7856 ground-truth objects. True positives dominate, detecting 7395 instances (80.08%). False positives represent a primary challenge, with 1379 instances. This pattern reveals the model's aggressive detection strategy, fearing over-detection to avoid missing objects. Localization Errors and Misclassifications are minimal, confirming accurate bounding box placement and clear class distinction when objects are detected. High recall affirms the model's effectiveness in detecting coffee trees, while moderate precision suggests potential for improvement through threshold tuning. The predominant False Positive category indicates that addressing over-detection, rather than missed objects, should be the primary focus for deployment optimization. Overall, the model is able to detect coffee plants well, as evidenced by the confusion matrix value in figure 3.

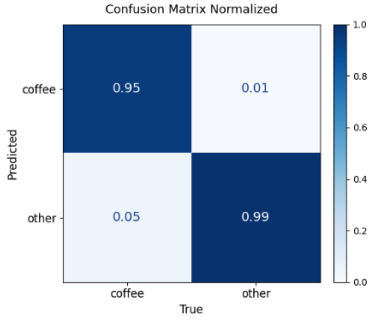

Figure 3. Confusion Matrix

B. *Discussion*

YOLOv8n's superiority (95.98% mAP@0.5) over larger variants reveals that coffee tree detection does not require extensive model capacity. YOLO8n is the lightest YOLOv8 variant, with 2 billion parameters. The relatively small amount of coffee field data results in a lighter and less complex computational process. The lightweight architecture and small amount of data are the reasons why YOLOv8n performs best. YOLOv8m, despite 8× more parameters and 8.4× larger size, achieves marginal improvement in mAP@0.5:0.95 while performing slightly worse on the primary metric. This pattern indicates that the relatively simple visual characteristics of coffee trees, consistent canopy appearance, structured layouts, and top-down viewpoints enable compact architectures to achieve near-optimal performance. This contrasts with natural image detection, where larger models consistently outperform smaller variants due to diverse object categories and complex occlusion patterns. Faster R-CNN's underperformance challenges the traditional view that two-stage detectors achieve superior accuracy. Several factors explain this result: YOLOv8's anchor-free detection and distribution focal loss provide more effective optimization for dense, small objects; the CSPDarknet backbone incorporates modern innovations absent in ResNet-50; and coffee tree detection lacks the extreme occlusions that traditionally favor two-stage refinement. Trees are largely visible from an aerial viewpoint with clear boundaries, making single-stage direct regression sufficient. EfficientDet's poor performance despite its acclaimed efficiency suggests that its architectural innovations do not compound at scale; BiFPN provides less benefit for agricultural imagery. Coffee trees exhibit relatively consistent size distribution compared to COCO's diverse multi-scale objects, reducing the value of compound scaling strategies. These findings emphasize that architectural

requirements vary substantially between domains, and benchmark performance does not guarantee transferability to agricultural applications.

The superiority of basic augmentation over both baseline and heavy augmentation reveals fundamental differences between structured aerial imagery and unstructured natural scenes. Basic augmentation succeeds by aligning with natural variation, simulating lighting conditions while maintaining the spatial structure of regular row spacing and predictable canopy patterns. Heavy augmentation fails because mosaic and mixup violate the plantation structure. Mosaic creates unrealistic spatial configurations with discontinuous tree patterns and incompatible lighting, introducing training samples that diverge from the actual data distribution. Mixup generates semi-transparent overlapping trees that never occur in reality. For aerial crop monitoring, simple geometric and photometric augmentation suffices, while advanced composition techniques introduce harmful artifacts. Practitioners should introduce augmentation incrementally, starting with basic transformations and adding complexity only if validation performance improves. Crops with structured layouts likely exhibit similar patterns, while less structured scenarios may benefit from more aggressive strategies.

Strong performance with limited data indicates coffee tree detection is relatively data-efficient compared to general object detection. Several factors contribute: visual consistency, clear figure-ground separation, limited intra-class variation, and structured spatial priors. Diminishing returns reveal that the model captures essential patterns from initial samples, with additional data primarily refining boundaries rather than introducing novel information. The 375-image subset contains representative samples that cover the main sources of variation (growth stages, lighting, canopy density, perspectives), with additional samples largely duplicating the learned patterns. In practice, these findings enable significant cost reductions.

Strong performance with limited data (94.94% mAP@0.5 with only 375 images, 98.76% of full performance) indicates coffee tree detection is relatively data-efficient compared to general object detection. Several factors contribute: visual consistency (uniform canopy appearance), clear figure-ground separation (distinct contrast against backgrounds), limited intra-class variation (bounded size range), and structured spatial priors (regular row spacing). Diminishing reveals that the model captures essential patterns from initial samples, with additional data primarily refining boundaries rather than introducing novel information. The 375-image subset contains representative samples that cover the main sources of variation (growth stages, lighting, canopy density, perspectives), with additional samples largely duplicating the learned patterns. In practice, these findings enable significant cost reductions. A model trained on our dataset could be fine-tuned for new regions, varieties, or tree crops with 100-200 region-specific images, enabling rapid deployment in data-scarce regions. By proving that the YOLOv8n variant has the best performance compared to other more complex variants, it indirectly allows this detection model to be implemented on devices such as drones. However, our findings reflect random sampling from a single dataset under consistent conditions. Generalization to fundamentally different contexts (different varieties, cultivation systems, geographical regions) may require more data. Cross-region validation would clarify whether 750 images suffice universally or represent a dataset-specific finding. Despite the advantages outlined, this study certainly has limitations. This limitation is that the model has not been tested on land data other than coffee, so implementation or testing on data other than coffee land is still not possible.

IV. CONCLUSION

This study conducted a comprehensive evaluation of deep learning architectures specifically for coffee tree detection. The performance comparison concluded that YOLOv8n was the optimal architecture, with a mAP@0.5 of 95.98%. This result demonstrates that one-stage architectures, such as YOLO, can outperform more complex models, confirming that the structured visual characteristics of plantations enable the use of lighter models without compromising performance. Furthermore, the study identified several practical applications of augmentation, indicating that moderate augmentation techniques are more effective than aggressive or extensive ones. In the data-efficiency comparison experiment, the detection model also achieved optimal performance with 750 images, suggesting diminishing returns beyond this point. The final experiment involved an error analysis, which revealed that the prevalence of false positives can be effectively mitigated by calibrating the confidence threshold without necessitating modifications to the architecture. Overall, this research contributes to bridging the gap between academic inquiry and practical application through a cost-effective strategy.

ACKNOWLEDGEMENT

The authors wish to express their gratitude to Institut Teknologi Sumatera for the research funding scheme titled “Penguatan Kelompok Keilmuan Institut Teknologi Sumatera” with grant number 2658aa/IT9.2.1/PT.01.03/2025.

REFERENCES

[1] S. Krishnan, “Sustainable Coffee Production,” *Oxford Research Encyclopedia of Environmental Science*, Jun. 2017, doi: https://doi.org/10.1093/acrefore/9780199389414.013.224

[2] X. Wang, C. Zhang, Z. Qiang, C. Liu, X. Wei, and F. Cheng, “*A Coffee Plant Counting Method Based on Dual-Channel NMS and YOLOv9 Leveraging UAV Multispectral Imaging*,” Remote Sensing, vol. 16, no. 20, pp. 3810–3810, Oct. 2024, doi: https://doi.org/10.3390/rs16203810

[3] J. Bolaños, Juan Carlos Corrales, and Liseth Viviana Campo, “Feasibility of Early Yield Prediction per Coffee Tree Based on Multispectral Aerial Imagery: Case of Arabica Coffee Crops in Cauca-

Colombia," Remote Sensing, vol. 15, no. 1, pp. 282–282, Jan. 2023, doi: https://doi.org/10.3390/rs15010282.
[4] F. Rovira-Más and Verónica Saiz-Rubio, "Crop Scouting and Surrounding Awareness for Specialty Crops," Agriculture automation and control, pp. 111–136, Jan. 2021, doi: https://doi.org/10.1007/978-3-030-70400-1_5.
[5] C. Bunn, P. Läderach, O. Ovalle Rivera, and D. Kirschke, "A bitter cup: climate change profile of global production of Arabica and Robusta coffee," Climatic Change, vol. 129, no. 1–2, pp. 89–101, Dec. 2014, doi: https://doi.org/10.1007/s10584-014-1306-x.
[6] A. F. Colaço and J. P. Molin, "Variable rate fertilization in citrus: a long term study," Precision Agriculture, vol. 18, no. 2, pp. 169–191, May 2016, doi: https://doi.org/10.1007/s11119-016-9454-9.
[7] W. H. Maes and K. Steppe, "Perspectives for Remote Sensing with Unmanned Aerial Vehicles in Precision Agriculture," Trends in Plant Science, vol. 24, no. 2, pp. 152–164, Feb. 2019, doi: https://doi.org/10.1016/j.tplants.2018.11.007.
[8] Y. Ampatzidis and V. Partel, "UAV-Based High Throughput Phenotyping in Citrus Utilizing Multispectral Imaging and Artificial Intelligence," Remote Sensing, vol. 11, no. 4, p. 410, Feb. 2019, doi: https://doi.org/10.3390/rs11040410.
[9] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, Apr. 2018, doi: https://doi.org/10.1016/j.compag.2018.02.016.
[10] L. Tang and G. Shao, "Drone remote sensing for forestry research and practices," Journal of Forestry Research, vol. 26, no. 4, pp. 791–797, Jun. 2015, doi: https://doi.org/10.1007/s11676-015-0088-y.
[11] L. S. Santana, G. A. e S. Ferraz, G. H. R. dos Santos, N. L. Bento, and R. de O. Faria, "Identification and Counting of Coffee Trees Based on Convolutional Neural Network Applied to RGB Images Obtained by RPA," Sustainability, vol. 15, no. 1, p. 820, Jan. 2023, doi: https://doi.org/10.3390/su15010820.
[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," openaccess.thecvf.com, 2014. https://openaccess.thecvf.com/content_cvpr_2014/html/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.html
[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Cv-foundation.org, pp. 779–788, 2016, Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html
[14] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," openaccess.thecvf.com, 2020. https://openaccess.thecvf.com/content_CVPR_2020/html/Tan_EfficientDet_Scalable_and_Efficient_Object_Detection_CVPR_2020_paper.html
[15] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, Jul. 2019, doi: https://doi.org/10.1186/s40537-019-0197-0.
[16] Z. Khan, Y. Shen, and H. Liu, "ObjectDetection in Agriculture: A Comprehensive Review of Methods, Applications, Challenges, and Future Directions," Agriculture, vol. 15, no. 13, p. 1351, Jun. 2025, doi: https://doi.org/10.3390/agriculture15131351.
[17] H. Raza, M. A. Bakr, S. D. Khan, Hira Batool, H. Ullah, and M. Ullah, "Benchmarking YOLO Models for Crop Growth and Weed Detection in Cotton Fields," AgriEngineering, vol. 7, no. 11, pp. 375–375, Nov. 2025, doi: https://doi.org/10.3390/agriengineering7110375.
[18] A. Kamilaris and Prenafeta-Boldú, Francesc X, "Disaster Monitoring using Unmanned Aerial Vehicles and Deep Learning," arXiv.org, 2024. https://arxiv.org/abs/1807.11805 (accessed Oct. 01, 2024).
[19] J. G. M. Esgario, R. A. Krohling, and J. A. Ventura, "Deep learning for classification and severity estimation of coffee leaf biotic stress," Computers and Electronics in Agriculture, vol. 169, p. 105162, Feb. 2020, doi: https://doi.org/10.1016/j.compag.2019.105162.
[20] E. Silva, J. B. Fragoso, Thuanne Paixão, A. B. Alvarez, and Facundo Palomino-Quispe, "A Low Computational Cost Deep Learning Approach for Localization and Classification of Diseases and Pests in Coffee Leaves," IEEE Access, pp. 1–1, Jan. 2025, doi: https://doi.org/10.1109/access.2025.3562832.
[21] M. A. Tamayo-Monsalve et al., "Coffee Maturity Classification Using Convolutional Neural Networks and Transfer Learning," IEEE Access, vol. 10, pp. 42971–42982, 2022, doi: https://doi.org/10.1109/access.2022.3166515.
[22] D. Su, H. Kong, Y. Qiao, and S. Sukkarieh, "Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics," Computers and Electronics in Agriculture, vol. 190, p. 106418, Nov. 2021, doi: https://doi.org/10.1016/j.compag.2021.106418.
[23] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," arXiv.org, 2019. https://arxiv.org/abs/1905.05055
[24] L. Liu et al., "Deep Learning for Generic Object Detection: A Survey," arXiv:1809.02165 [cs], Aug. 2019, Available: https://arxiv.org/abs/1809.02165
[25] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893, 2005, doi: https://doi.org/10.1109/cvpr.2005.177.
[26] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, Sep. 2010, doi: https://doi.org/10.1109/tpami.2009.167.
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105, 2012, Available: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
[28] W. Liu et al., "SSD: Single Shot MultiBox Detector," Computer Vision – ECCV 2016, vol. 9905, no. 5, pp. 21–37, 2016, doi: https://doi.org/10.1007/978-3-319-46448-0_2.
[29] R. Girshick, "Fast R-CNN," openaccess.thecvf.com, 2015. https://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html
[30] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," openaccess.thecvf.com, 2017. https://openaccess.thecvf.com/content_cvpr_2017/html/Redmon_YOLO9000_Better_Faster_CVPR_2017_paper.html
[31] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv, vol. 1, Apr. 2020, Available: https://arxiv.org/abs/2004.10934
[32] R. Varghese and Sambath M, "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," IEEE, Apr. 2024, doi: https://doi.org/10.1109/adics58448.2024.10533619.