CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2026.048

An efficient and lightweight YOLOv8-based model with multi-scale texture representation for durian leaf disease classification

Van Quyet Nguyen[1], Quang Hung Ha[2], Truong Trinh Phung[2], Minh Trung Vu[1], Ha Duc Chu[1], and Minh Trien Pham[3*]

[1]Faculty of Agricultural Technology, VNU University of Engineering and Technology, Viet Nam
[2]Faculty of Information Technology, VNU University of Engineering and Technology, Viet Nam
[3]Faculty of Electronics Engineering 1, Posts and Telecommunications Institute of Technology, Viet Nam
*Corresponding author (trienpm@ptit.edu.vn)

Article info.

Received 21 Jan 2026
Revised 5 Jun 2026
Accepted 29 May 2026

Keywords

Durian, leaf disease, YOLO, deep learning, PartialTriDW

ABSTRACT

Durian (Durio zibethinus) is a high-value fruit crop in Viet Nam, with rapidly increasing production, which underscores the need for accurate leaf condition recognition in real orchards. Field images are challenging because symptoms vary in scale and are captured under uncontrolled illumination and complex backgrounds, while many deep models improve accuracy at the cost of higher computation. This study proposes an efficient YOLOv8-based classifier for durian leaf images. The method removes the deepest downsampling stage to preserve higher-resolution features and introduces PartialTriDW, a lightweight tri-branch depthwise block with adaptive softmax fusion and low-cost channel mixing, to enhance multi-scale texture representation in the early and mid-backbone stages. Experiments on a five-class dataset of 4,437 field images, trained from scratch, show that the YOLOv8 baseline achieves 0.918 accuracy, the CutP5 variant achieves 0.927 accuracy, and the proposed model reaches 0.949 accuracy with 0.56 million parameters and 3.3 GFLOPs, while maintaining comparable latency. The results indicate that preserving spatial detail and strengthening multi-scale texture modeling improve practical durian leaf classification with controlled computational cost.

1. INTRODUCTION

Durian (*Durio zibethinus*) is a high-value tropical fruit crop whose production has expanded rapidly in Viet Nam in recent years. In the first nine months of 2025, the national statistics office of Viet Nam reported durian output reached 636.3 thousand tons, corresponding to a 19.3 per cent year-on-year increase, underscoring both the scale and the accelerating economic importance of the crop (The Ministry of Finance of the Socialist Republic of Viet Nam, 2025). As production intensifies, maintaining canopy health becomes increasingly consequential because leaf disorders can impair photosynthetic capacity, weaken tree vigor, and ultimately reduce marketable yield and fruit quality. In practice, durian leaf health management continues to rely heavily on field scouting and visual inspection, often supplemented by expert consultation and chemical control measures. While these approaches remain indispensable, they are constrained by labor demand, subjectivity, and difficulty in consistently recognizing early or ambiguous symptoms under variable field conditions. Prior reviews of leaf disease recognition emphasize that manual diagnosis can be unreliable when symptoms are diverse, visually similar across conditions, or observed by non-specialists, which may lead to delayed or inappropriate treatment decisions and excessive pesticide use (Zhao et al., 2025). These limitations are particularly relevant for orchard imagery captured in real field settings, where lighting changes, background clutter, and viewpoint variation frequently obscure the most diagnostic leaf cues.

Machine learning, and deep learning in particular, provides a pathway toward scalable and consistent image-based diagnosis by learning discriminative patterns directly from data. For durian leaves, recent public resources and studies have begun to support automated recognition, including datasets and classification pipelines targeting common leaf conditions such as algal symptoms, phomopsis-related spots, and blight-like manifestations (Thanh et al., 2025). Nevertheless, the same work highlights persistent gaps in the durian literature, with many prior studies limited in sample size and class diversity, reducing robustness in practical deployment (Thanh et al., 2025). Beyond data scale, a recurring technical difficulty is feature representation during field acquisition: a model must remain sensitive to fine-grained lesion texture and edge-level detail while also handling broad diseased patches and nuisance variations caused by illumination and background.

This study targets durian leaf image classification under real orchard acquisition conditions, where diagnostically relevant cues can be subtle and appear at multiple spatial scales. This study adopts the Ultralytics YOLOv8 classification framework as the baseline for three reasons. First, its CSPDarknet backbone with C2f modules is inherently multi-scale, progressively extracting both fine-grained texture and broader contextual features across successive network stages, which directly aligns with the challenge of detecting subtle symptom cues at varying scales in orchard imagery (Jocher et al., 2023). Second, prior work has demonstrated the competitiveness of the YOLOv8 classification backbone on plant leaf disease tasks, including citrus leaf disease classification (Zhu et al., 2025), general plant disease classification across multiple crop datasets (Balaha, 2026), and cotton leaf disease classification using field-captured images achieving top-1 accuracy of 99.60% (Joshi et al., 2025), establishing it as a validated baseline for this problem domain. Third, the Ultralytics framework provides a unified, reproducible training and evaluation pipeline, ensuring that observed performance gains are attributable to the proposed architectural modifications rather than to implementation variance. In this paper, the study does not rely on specialized data augmentation to compensate for field variability; instead, the study emphasizes architectural design choices that preserve and strengthen fine symptom evidence within the early backbone feature hierarchy, so that the model can learn robust representations directly from the available data.

The central architectural hypothesis is that, for field-captured durian leaves, excessively deep and highly downsampled representations can become overly coarse and may underrepresent lesion morphology. Standard YOLO-style backbones progressively reduce spatial resolution through stages commonly denoted P2 to P5; however, when the input is relatively small, the deepest stage can collapse spatial structure into a very small grid, encouraging reliance on global context rather than fine-grained detail. Accordingly, this study removes the P5 stage from the YOLOv8 classification backbone to retain a higher-resolution terminal representation and reduce unnecessary computation for this application.

To strengthen early-stage representation without incurring significant computational overhead, the paper introduces PartialTriDW, an efficient multi-scale feature-mixing block integrated into the shallow stages. PartialTriDW performs feature splitting followed by three parallel depthwise convolution branches with kernel configurations 3×3, 5×5, and 3×3 with dilation, and then applies softmax-based fusion to adaptively weight these branches per image, in the spirit of dynamic receptive field selection mechanisms (Li et al., 2019). Depthwise convolution is adopted for efficiency, consistent with established lightweight design principles (Howard et al., 2017). After multi-scale fusion, PartialTriDW employs channel shuffle and grouped 1×1 mixing to promote cross-channel information exchange at low cost, following the rationale introduced in ShuffleNet (Zhang et al., 2018). Finally, the block incorporates a residual layer scale mechanism to stabilize optimization when training from scratch, aligning with prior evidence that LayerScale improves training stability in deep residual style architectures (Touvron et al., 2021).

PartialTriDW is integrated into the P2, P3, and P4 stages of YOLOv8 classification, thereby prioritizing sensitivity to fine textures and medium-scale diseased regions while keeping computational cost controlled. This design is particularly suited to real field durian leaf imagery, where discriminative cues may appear as both small, localized textures and larger irregular patches. The resulting model is intended to provide a practical improvement for durian leaf condition recognition under deployment-like constraints, especially in settings where pretrained weights are unavailable or impractical to obtain due to connectivity or regulatory constraints, and where models must be trained entirely from locally collected field data.

2. MATERIALS AND METHODS

2.1. Dataset

This work uses a publicly available dataset of durian leaves, released for an image classification task. The dataset was collected from durian farms in the Southeast region of Viet Nam with the explicit intention of reflecting real orchard observation conditions, where images are captured under uncontrolled illumination, complex backgrounds, and heterogeneous symptom presentation (Thang, 2025). In particular, the acquisition procedure relied on mobile phone photography of leaves directly on the tree, with images taken under uneven lighting, variable weather conditions such as sunny conditions and light rain, and diverse shooting angles and capture distances, as shown in Figure 1.

Figure 1. Overview of the durian leaf dataset by class

(a) Phomopsis leaf spot, (b) Leaf blight, (c) Healthy leaf, (d) Allocaridara attack, and (e) Algal leaf spot.

The dataset includes five categories, as shown in Figure 2, corresponding to common durian leaf conditions: algal leaf spot, allocaridara attack, healthy leaf, leaf blight, and phomopsis leaf spot. After screening for label consistency, the final dataset contains 4,437 images. All images were resized to 224x224 pixels and stored in JPG format to standardize model input and enable reproducible experimentation. Ground truth labels were produced through an expert-driven protocol. Two agricultural experts independently annotated the dataset after aligning on class definitions and representative exemplars. The dataset documentation reports a Cohen's kappa of 98.25%, and samples with expert disagreement were excluded to reduce label noise and avoid ambiguous supervision. For model development and evaluation, the dataset was partitioned into training, validation, and test as shown in Table 1.

Table 1. Summary of the dataset partitioning

Split	Ratio	Images
Train	80%	3547
Val	10%	441
Test	10%	449

2.2. Experimental setup and model selection

The experiments were conducted in Visual Studio Code on a workstation equipped with an NVIDIA Quadro RTX 5000 GPU, 128 GB RAM, and an Intel® Xeon® Silver 4216 CPU. All models in this study are trained using SGD with momentum 0.937 and weight decay 5×10^{-4}. The initial learning rate was $lr_0 = 0.01$, decayed via a cosine schedule to a minimum of 1×10^{-4} over 100 epochs (batch size 16, input size 224×224, 8 warm-up epochs, automatic mixed precision enabled).

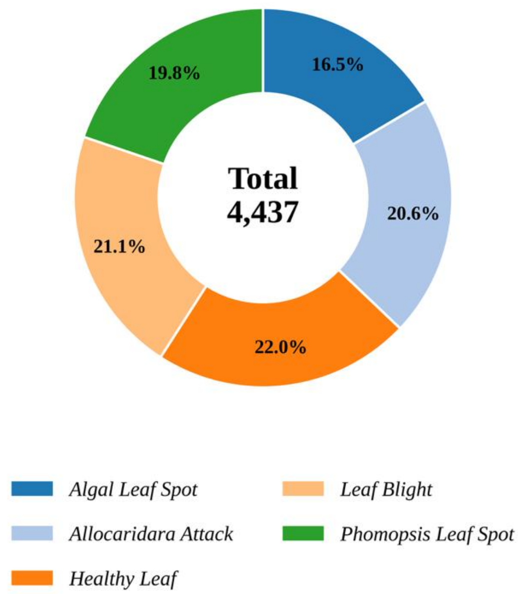

Figure 2. Class distribution of the dataset

Inference latency and throughput were measured on the same hardware under a controlled protocol: each model received single-sample inputs (batch size 1), was preceded by ten warm-up passes to reach a stable GPU state, and was then timed over 100 consecutive forward passes using GPU-side event timing. The reported latency is the mean over these runs; variance across runs was below 0.5 ms for all models. Throughput, in frames per second, was derived as the reciprocal of the mean latency.

Based on the model results in Table 2, YOLOv8 was selected as the baseline model because it offers the most suitable balance between classification accuracy and inference efficiency. YOLOv8 achieved a higher accuracy of 0.918 while maintaining low computational complexity at 3.4 GFLOPs and a smaller parameter count of 1.44M. YOLO11 (Jocher & Qui, 2024) reported a similar complexity of 3.3 GFLOPs but delivered lower accuracy at 0.912, so it did not meet the performance objective under comparable compute. YOLO12 (Tian et al., 2025) achieved an accuracy of 0.907 but had 0.17s latency and only 5.85 FPS, which does not meet practical runtime constraints. Therefore, YOLOv8 was adopted as the reference architecture for the proposed modifications and the ablation studies.

Table 2. Model selection results

Model	Params(M)	FLOPs(GFLOPs)	Accuracy	Latency(s)	FPS
YOLOv8	**1.44**	3.4	**0.918**	0.152	**6.55**
YOLO11	1.54	**3.3**	0.912	0.158	6.32
YOLO12	1.72	3.6	0.907	0.171	5.85

2.3. Methodology

2.3.1. PartialTriDW block

To enhance multi-scale texture modeling in shallow and mid-level feature stages while keeping computation low, this study introduces the PartialTriDW block as an efficient module that combines multi-scale spatial filtering, adaptive branch fusion, lightweight channel mixing, and stable residual learning (Figure 3). The input feature map is $x \in R^{B\times C\times H\times W}$, where the block first aligns the channel dimension when necessary. If the incoming channels differ from the target channels, a pointwise projection is applied:

$$x \leftarrow Conv_{1\times1}(x) \quad (1)$$

The block then supports a channel partition mechanism controlled by a ratio p, which determines how many channels are processed and how many are preserved as a bypass. In this paper, the full channel-processed set is used, so that multi-scale re-encoding is applied to the entire representation rather than to a small subset. Denoting the processed tensor by x_{proc}, and $x_{proc} = x$ under full processing.

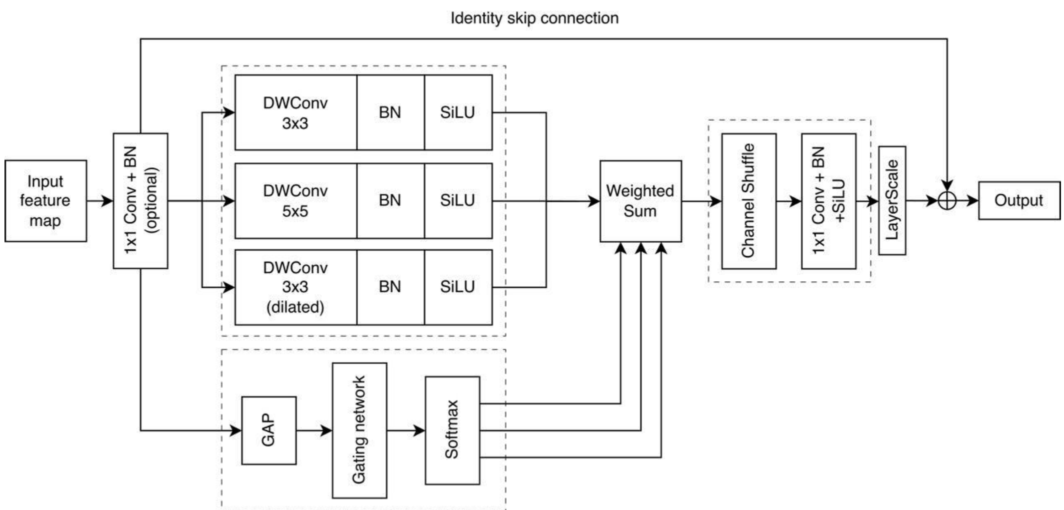

Figure 3. Structure of the proposed PartialTriDW block

The core of PartialTriDW is a tri-scale spatial transform implemented with three parallel depthwise convolution branches. Depthwise convolution applies one spatial filter per channel, which is cheaper than standard convolution when the channel count is large. The three branches are shown in Equation 2:

$$\begin{gathered} b_1 = DWConv_{3\times3}(x), \\ b_2 = DWConv_{5\times5}(x), \\ b_3 = DWConv_{3\times3,\ d=2}(x) \end{gathered} \quad (2)$$

The 3×3 branch focuses on very local texture changes and sharp boundaries, which helps when the symptoms are small and detailed. The 5×5 branch aggregates information over a wider neighborhood, which is more suitable when symptoms cover larger regions. The dilated 3×3 branch increases the effective receptive field without reducing the feature map resolution, so the block can capture broader context while preserving fine spatial detail.

Instead of always averaging them or always summing them equally, PartialTriDW computes image-dependent fusion weights. This is done by first compressing the processed feature map into a global descriptor using global average pooling (GAP), which produces a compact summary of the current image features:

$$s = GAP(x_{proc}) \in R^{B\times C\times 1\times 1} \qquad (3)$$

This descriptor is then passed through a lightweight gating network to produce three branch scores, one score per branch. The gating network consists of two sequential 1×1 convolutions separated by Batch Normalization (BN) and a SiLU activation. Specifically, given the global descriptor $s \in R^{B\times C\times 1\times 1}$, the network computes in Equation 4:

$$\begin{aligned} h &= SiLU\left(BN\left(Conv_{\{1\times1\}(s;\, C\to H)}\right)\right) \\ H &= \max\left(\frac{C}{4}, 16\right) \\ Gate(s) &= Conv_{\{1\times1\}(h;\, H\to 3)} \end{aligned} \qquad (4)$$

where H is the hidden dimension, C is the number of input channels, and $Conv_{\{1\times1\}(\cdot;\, a\to b)}$ denotes a 1×1 convolution projecting from a to b channels. This gating pathway adds only a small number of parameters (two 1×1 convolution layers with $C\times H + H\times 3$ weights in total). Softmax is then applied to the Gate(s) to produce three positive weights w_1, w_2, w_3 that sum to one as Equation 5:

$$[w_1, w_2, w_3] = softmax(Gate(s)), \quad \sum_{i=1}^{3} w_i = 1 \qquad (5)$$

The fused feature map is then computed as a weighted sum, shown in Equation 6:

$$\mathrm{y} = w_1 b_1 + w_2 b_2 + w_3 b_3 \qquad (6)$$

For an input dominated by small speckles, the model can assign a larger weight to the 3×3 branch. For an input dominated by larger blotches, the model can shift weight toward the 5×5 branch or the dilated branch. In other words, the block learns to choose the appropriate scale for each image, a concept closely related to adaptive receptive field ideas in multi-branch architectures.

After multi-scale fusion, the block improves channel interaction using two inexpensive steps. Firstly, it applies channel shuffle, which is simply a permutation that mixes channels across groups. Channel shuffle is widely used to prevent channels from becoming isolated when group-structured operations are involved. Secondly, the block applies pointwise mixing using a 1×1 convolution. This step recombines the fused information across channels to produce a stronger representation as Equation 7:

$$u = Conv_{1\times1}(Shuffle(y)) \qquad (7)$$

Finally, PartialTriDW uses a residual connection with a learnable LayerScale coefficient to stabilize training from scratch. The output is shown in Equation 8 as:

$$out = x + \gamma u \qquad (8)$$

The scale parameter γ is initialized near zero, so at the beginning of training, the block behaves almost like the identity mapping. As learning progresses, γ increases and the block contributes more strongly. This strategy improves optimization stability when inserting new modules and has been reported to help training in deep architectures.

2.3.2. Integration of PartialTriDW into YOLOv8 classification backbone

Figure 4 illustrates how the proposed PartialTriDW block is incorporated into the YOLOv8 classification architecture. In the overall pipeline, the image is processed by a stage-wise backbone that progressively downsamples the feature map and refines representations before forwarding the final embedding to the classification head. PartialTriDW is inserted directly into the backbone at the shallow and intermediate stages, corresponding to P2, P3, and P4 in the conventional feature hierarchy. As shown in Figure 4, each insertion point follows the standard stage transition pattern: a downsampling convolution produces the next resolution level, and PartialTriDW then enhances the resulting features before the subsequent backbone refinement blocks propagate the representation forward. This placement ensures that the proposed module operates on feature maps that still preserve sufficient spatial detail to capture symptom morphology, while also being exposed to progressively richer contextual information across stages.

The motivation for integrating PartialTriDW at these stages is grounded in the visual properties of the field captured by durian leaf images.

Discriminative evidence for leaf conditions often appears as subtle texture changes, small speckles, irregular lesion boundaries, or broad discolored regions. These cues are most reliably represented at higher spatial resolutions. In classification settings with an input size of 224×224, repeated downsampling can rapidly compress the feature map, weakening fine-grained lesion structure and increasing the risk that the model relies on coarse global context. By enhancing features at P2 to P4, the architecture strengthens symptom-relevant evidence early in the hierarchy, so that later layers receive representations that already encode multi-scale lesion patterns rather than needing to recover them from overly compressed features.

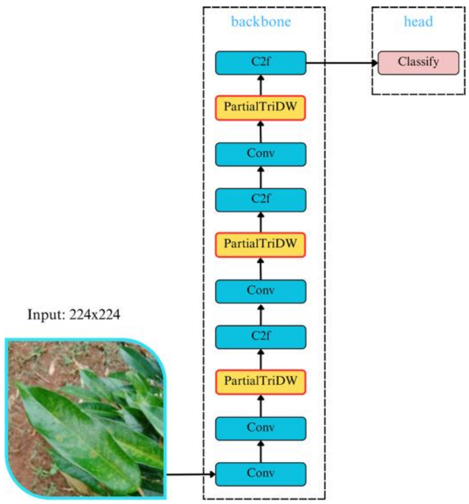

Figure 4. Architecture of the proposed YOLOv8-based classifier

The role of PartialTriDW within this backbone is to provide multi-scale texture modeling without incurring the computational cost associated with large channel expansion. As detailed in Figure 4, the block applies three parallel depthwise convolution branches with complementary receptive fields, and it fuses them using adaptive softmax weights derived from a lightweight gating pathway. This design is particularly appropriate for orchard imagery because the symptom scale varies with disease type, capture distance, and viewpoint. The adaptive fusion allows the network to increase the contribution of the most relevant branch for each image, so that small, speckled patterns and larger blighted regions can be emphasized when present, while still maintaining a unified backbone structure.

The integration strategy also preserves stable optimization under training from scratch. PartialTriDW includes a residual formulation with a learnable LayerScale parameter initialized near zero, which makes the block behave close to an identity mapping in early training. This property reduces the risk that newly introduced modules disturb gradient flow or cause unstable updates at the beginning of optimization. As training progresses, the LayerScale parameter increases and the module contributes more strongly, enabling the backbone to learn improved representations while maintaining the baseline training dynamics of YOLOv8 classification.

3. RESULTS AND DISCUSSION

3.1. Results

The quantitative results for the baseline model, baseline model removed P5, and the proposed model are summarized in Table 3. All models were trained under the same experimental setting, and the reported metrics include classification accuracy, model size, and inference efficiency. The baseline YOLOv8 classifier achieved an accuracy of 0.918 with 1.44M parameters and 3.4 GFLOPs. Under the same evaluation protocol, the measured inference latency was 0.152s, and the corresponding throughput was 6.55 FPS. These results provide the reference point for assessing the impact of modifying backbone depth and enhancing feature extraction in earlier stages.

When the deepest backbone stage was removed, the resulting YOLOv8 CutP5 variant achieved higher accuracy than the baseline. Accuracy increased from 0.918 to 0.927, corresponding to an absolute gain of 0.009. Model size was reduced substantially, with parameters decreasing from 1.44M to 0.53M. Computational cost also decreased from 3.4 GFLOPs to 3.0 GFLOPs. Runtime measurements followed the same trend. Latency decreased from 0.152s to 0.139s, and throughput increased from 6.55 FPS to 7.21 FPS. These results show that removing the P5 stage improved both predictive performance and efficiency under the current input resolution and dataset characteristics.

The improved YOLOv8 model obtained the strongest classification performance among the evaluated variants. The model achieved an accuracy of 0.949, corresponding to an absolute improvement of 0.030 over the baseline and 0.021 over the CutP5 variant. The improved model also remained compact. It used 0.56M parameters, which is far below the baseline model size while being slightly higher than the CutP5 variant. The computational cost was 3.3 GFLOPs, close to the baseline and higher than YOLOv8-CutP5. Inference efficiency remained similar to the baseline. The measured latency was 0.151s, and the throughput was 6.62 FPS. These results indicate that the accuracy gains of the improved model were not accompanied by a substantial increase in computational burden, and that the runtime profile remained within the same operating range as the baseline model.

Table 3. Performance of YOLOv8 variants on the durian leaf dataset

Model	Params(M)	FLOPs(GFLOPs)	Accuracy	Latency(s)	FPS
YOLOv8 base	1.44	3.4	0.918	0.152	6.55
YOLOv8-CutP5	**0.53**	**3.0**	0.927	**0.139**	**7.21**
Improved YOLOv8	0.56	3.3	**0.949**	0.151	6.62

Table 4. Per-class evaluation of the improved YOLOv8 on the test set

Class	Precision	Recall	F1-score
Algal leaf spot	0.944	0.905	0.924
Allocaridara attack	0.946	0.957	0.951
Healthy leaf	0.925	**1.000**	0.961
Leaf blight	**0.978**	0.947	**0.963**
Phomopsis leaf spot	0.954	0.921	0.937
Average	0.949	0.946	0.947

Table 4 reports per-class precision, recall, and F1-score for the proposed model on the test set. Performance is consistent across all five classes, with F1-scores ranging from 0.924 to 0.963. Leaf blight achieves the highest precision (0.978) and F1-score (0.963), reflecting its relatively distinct visual pattern. Healthy leaf achieves perfect recall (1.000, F1 = 0.961), indicating that the model reliably identifies the absence of disease symptoms. Allocaridara attack achieves a precision of 0.946, a recall of 0.957, and an F1-score of 0.951, showing well-balanced performance. Phomopsis leaf spot achieves a precision of 0.954 and a recall of 0.921, yielding an F1-score of 0.937. Algal leaf spot records the lowest F1-score (0.924), with a precision of 0.944 and recall of 0.905, consistent with the inter-class confusion patterns identified in the qualitative analysis. The macro-averaged precision, recall, and F1-score are 0.949, 0.946, and 0.947, respectively, indicating balanced performance across classes despite the moderate class imbalance in the test set.

The results in Table 3 establish three main observations. Firstly, backbone truncation alone produced a consistent improvement in accuracy while simultaneously reducing parameters, GFLOPs, and latency. Secondly, the improved model achieved the largest accuracy gain while keeping its size substantially lower than the baseline. Thirdly, the improved model maintained baseline runtime performance while improving accuracy by a meaningful margin.

Table 5 extends the comparison to three representative lightweight classifiers trained under the identical experimental protocol. MobileNetV3-Small, with 1.52M parameters, achieves a test accuracy of 0.927. ShuffleNetV2-x1.0, with 1.26M parameters, achieves 0.938. Both models have more parameters than the proposed model (0.56M) yet yield lower accuracy. EfficientNet-B0 achieves the highest accuracy of 0.958 with 4.01M parameters. Among all architectures with fewer than 1M parameters, the proposed model achieves the highest accuracy in this comparison.

Table 5. Comparison with representative lightweight classifiers

Model	Params(M)	Accuracy
MobileNetV3-Small	1.52	0.927
ShuffleNetV2	1.26	0.938
EfficientNet-B0	4.01	**0.958**
Improved YOLOv8	**0.56**	0.949

Table 6 presents an ablation study isolating the contribution of each component. Removing adaptive fusion produces the largest accuracy reduction, from 0.949 to 0.930, indicating that image-dependent branch weighting is the most consequential component of PartialTriDW. Removing the dilated branch reduces accuracy to 0.943, confirming that extended receptive field coverage contributes to multi-scale representation beyond what the 3×3 and 5×5 branches provide. Removing LayerScale leads to a smaller decrease to 0.946, consistent with its role in stabilizing training from scratch rather than directly shaping feature representations. Retaining the P5 stage alongside

PartialTriDW yields a test accuracy of 0.950, the highest among all ablation variants, but at a parameter count of 1.48M, approximately three times that of the proposed model. This result confirms that P5 removal is necessary to maintain the efficiency profile of the proposed architecture. Replacing PartialTriDW with the plain CutP5 backbone yields 0.927, indicating that the accuracy gain over YOLOv8-CutP5 is attributable to the PartialTriDW blocks rather than backbone truncation alone.

The confusion matrices in Figure 5 complement the ablation results. The full model (panel a) distributes prediction errors across class pairs without a dominant off-diagonal cluster, indicating no systematic class-specific failure. Removing adaptive fusion (panel d) increases confusion between algal leaf spot and allocaridara attack relative to the full model, consistent with the accuracy drop in Table 6 and reflecting the difficulty of separating these classes without image-dependent scale selection. Removing the dilated branch (panel c) produces a similar pattern at a smaller magnitude. Panels (e) and (f) show confusion distributions close to the full model, in line with the smaller accuracy differences reported for those variants.

Table 6. Ablation study on PartialTriDW components

Model	Params(M)	GFLOPs	Accuracy
Improved model	0.56	3.3	**0.949**
Improved model without a dilated branch	0.56	3.3	0.943
Improved model without adaptive fusion	0.56	3.3	0.930
Improved model without LayerScale	0.56	3.4	0.946
YOLOv8 base with PartialTriDW	1.48	3.7	0.950
YOLOv8 without P5 layer and PartialTriDW	**0.53**	**3.0**	0.927

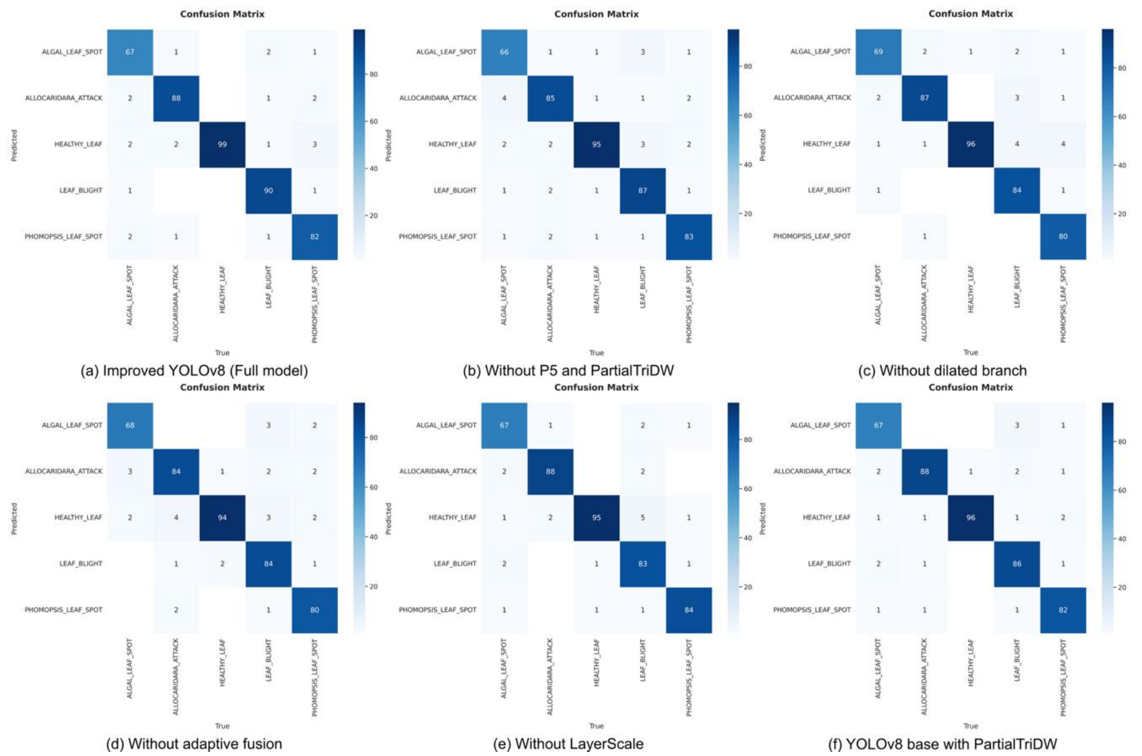

Figure 5. Confusion matrices for the improved model and the ablation variants evaluated on the test set

(a) Improved YOLOv8 (full model); (b) without P5 and PartialTriDW; (c) without dilated branch; (d) without adaptive fusion; (e) without LayerScale; (f) YOLOv8 base with PartialTriDW only (no P5 removal). Values represent raw sample counts per cell.

3.2. Discussion

The results in Table 3 indicate that durian leaf classification in real orchard imagery is strongly affected by the preservation of spatial detail and texture cues across the backbone hierarchy. Two consistent patterns emerge: removing the deepest downsampling stage improves both accuracy and efficiency, and strengthening feature extraction in early and mid-stages produces the largest accuracy gain while keeping model size and runtime within a practical range. The improvement observed for YOLOv8 CutP5 suggests that the deepest downsampling stage is not beneficial for this task at the current input resolution. At 224×224, repeated stride-based reduction can compress symptom evidence into a very small spatial grid. Fine lesion boundaries, sparse speckles, and subtle texture transitions become harder to represent after aggressive downsampling. Under these conditions, the representation can drift toward global cues that are not consistently related to the target class, including background patterns and illumination artifacts. The CutP5 variant preserves a higher resolution terminal representation. This provides a plausible explanation for the increase in accuracy, along with fewer parameters, lower GFLOPs, and faster inference. The following sections examine the contribution of each PartialTriDW component and the behavior of the adaptive gating mechanism across disease classes.

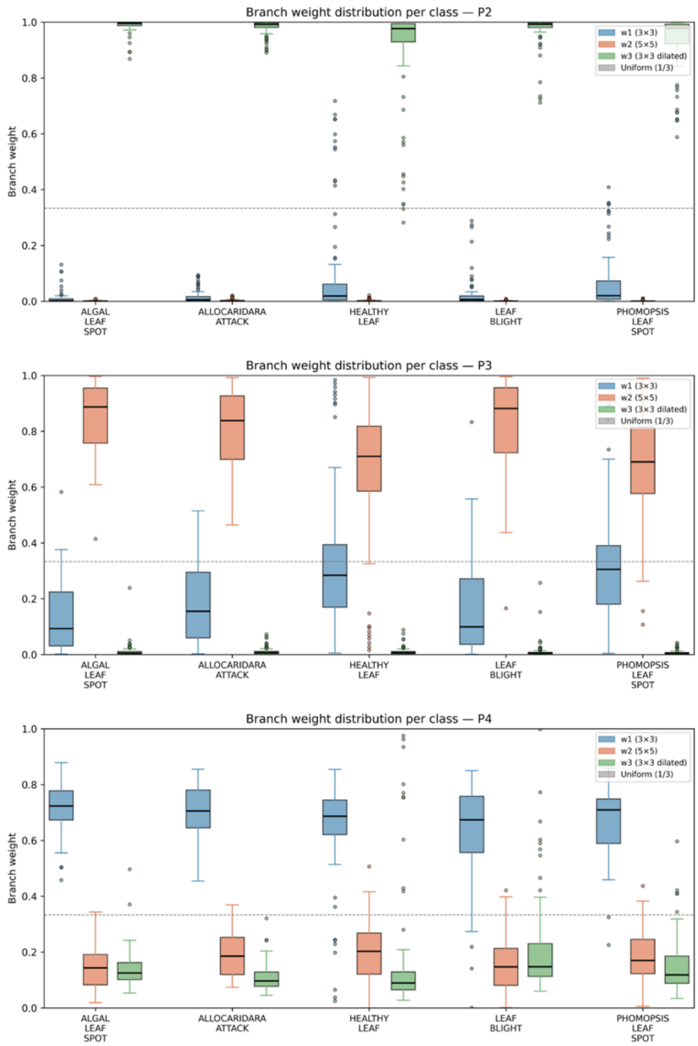

Figure 6. Learned branch weight distributions of the PartialTriDW gating network across the test set, stratified by disease class and backbone stage (P2, P3, P4)

The dashed line indicates uniform weighting (1/3) for reference.

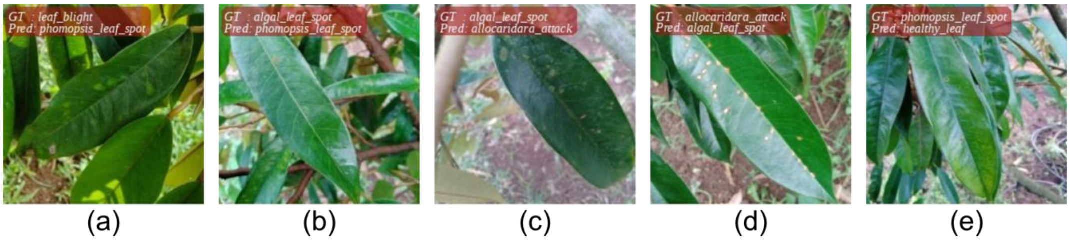

Figure 7. Representative failure cases of the proposed model

Each panel shows the ground-truth label (GT) and the predicted label (Pred). From left to right: (a) leaf blight misclassified as phomopsis leaf spot; (b) algal leaf spot misclassified as phomopsis leaf spot; (c) algal leaf spot misclassified as allocaridara attack; (d) allocaridara attack misclassified as algal leaf spot; (e) phomopsis leaf spot misclassified as healthy leaf.

The ablation results in Table 6 clarify the contribution of each architectural component. Removing adaptive softmax fusion reduces accuracy from 0.949 to 0.930, the largest single drop across all variants. This result supports the design premise that image-dependent branch weighting is necessary to handle the scale variability in field-captured durian leaf imagery, as uniform aggregation of branch outputs does not adequately adapt to the range of symptom scales present within and across classes. Removing the dilated branch produces a moderate reduction to 0.943, confirming that extended receptive field coverage provides complementary information beyond what the 3×3 and 5×5 branches capture. The LayerScale ablation yields a smaller decrease to 0.946, consistent with its stabilizing rather than representational role. Taken together, all three components contribute positively, with adaptive fusion accounting for the largest share of the gain in accuracy. The confusion matrices in Figure 5 support this interpretation: panel (d), corresponding to the variant without adaptive fusion, shows the most pronounced increase in off-diagonal counts, particularly for the algal leaf spot and allocaridara attack pair, which are the classes most sensitive to symptom scale.

The proposed architecture yields a larger improvement than truncation alone, indicating that preserving resolution is necessary but not sufficient. The strongest gain is consistent with the need for richer texture modeling at stages where symptom morphology is still explicit. Multi-scale filtering supports this requirement because leaf symptoms occur across a range of sizes within and across classes. Small spots and early-stage symptoms demand localized receptive fields. Larger necrotic or blighted regions require a broader context. Adaptive branch fusion provides a mechanism that can shift emphasis across these scales at the image level. The efficiency trends in Table 3 show that the accuracy gain is achieved without a proportional increase in computational cost, which supports the design objective of improving representational strength without relying on heavy channel expansion.

To further assess whether the adaptive gating mechanism operates in accordance with the design motivation, branch weight distributions were extracted across the full test set and stratified by disease class and backbone stage. The results are shown in Figure 6. At P2, the dilated 3×3 branch (w3) receives near-exclusive weight across all five classes (mean 0.90-0.99), indicating that the shallowest stage relies on expanded receptive fields to encode leaf-level spatial context. At P3, the 5×5 branch (w2) becomes consistently dominant (mean 0.66-0.86). Disease classes with broader symptom distributions, including algal leaf spot (0.857) and leaf blight (0.825), receive higher weights on this branch than classes with more localized or ambiguous presentations, such as healthy leaf (0.662) and phomopsis leaf spot (0.676). At P4, the 3×3 branch (w1) dominates uniformly across all classes (mean 0.63-0.72), suggesting that the deeper stage refocuses on local texture detail for final discriminative encoding. This progressive shift from dilated context at P2, through medium-scale aggregation at P3, to local refinement at P4 is consistent with the multi-scale design rationale of PartialTriDW, and indicates that the gating mechanism learns stage-appropriate scale preferences rather than converging to uniform weighting.

The results in Table 5 position the proposed model relative to established lightweight classifiers. Among architectures with fewer than 1M parameters, the proposed model achieves the highest accuracy in this comparison, outperforming both MobileNetV3-Small and ShuffleNetV2-x1.0 while using fewer parameters than either. EfficientNet-B0 achieves higher accuracy by approximately one percentage point, attributable to its larger parameter count (4.01M) and compound-scaled design. All classifiers in Table 5 were trained under the identical from-scratch protocol to ensure a fair comparison. Under the additional constraint of training from scratch without pretrained representations, this result indicates that the proposed architectural choices provide an effective inductive bias for this classification task within a constrained parameter budget.

The findings align with a broader understanding of visual recognition and plant disease analysis that multi-scale feature extraction and preservation of spatial detail are important when discriminative cues are texture-dominated and spatially localized. Similar motivations appear in prior work on adaptive receptive field selection and efficient depthwise-based modeling for visual tasks, where strong performance can be obtained by combining lightweight spatial operators with selective feature integration rather than increasing backbone depth alone. At the same time, the magnitude of the benefit depends on the dataset, image resolution, and symptom characteristics, so direct transfer of the numerical gains to other settings should not be assumed.

Qualitative error inspection provides additional context. Representative failure cases are shown in Figure 7. The most frequent confusion pairs involve visually similar classes. Algal leaf spot and allocaridara attack are mutually confused (cases c and d), likely because both conditions produce small, scattered marks that are difficult to separate under variable illumination and at the capture distances typical of mobile phone photography. Confusions between algal leaf spot and phomopsis leaf spot (case b), and between leaf blight and phomopsis leaf spot (case a), reflect overlapping symptom morphology: both pairs share irregular discoloration patterns that differ primarily in texture detail and lesion boundary sharpness. Case (e), where phomopsis leaf spot is predicted as a healthy leaf, occurs when symptom evidence is sparse and occupies a small fraction of the visible leaf area. These observations suggest that remaining errors are driven by inter-class visual similarity and limited symptom saliency under uncontrolled field conditions, rather than by insufficient model capacity.

Several limitations should be acknowledged. Evaluation is conducted on a single public dataset and a fixed input size, so generalization across orchards, seasons, devices, and management practices remains uncertain. All models in this study were trained from scratch to reflect a deployment scenario without access to external pretrained resources; a systematic comparison with ImageNet-pretrained variants of the same architectures remains an open direction for future work under less constrained conditions. All results reported in this study are based on a single fixed 80/10/10 data split; multi-seed evaluation to quantify variance across random initializations was not completed within the current revision scope and remains an important methodological direction for future work. Latency and throughput depend on the specific hardware and software environment used for measurement, so the reported values should be interpreted as comparative results under a consistent setting. Despite these constraints, the results support the central conclusion that reducing over-downsampling and enhancing multi-scale texture modeling at early and mid-backbone stages can improve practical durian leaf classification while maintaining an efficient inference profile. Cross-domain evaluation using images collected from additional orchards, capture devices, and seasons remains the most critical next step toward establishing the broader applicability of the proposed approach.

4. CONCLUSION

The study provides evidence that practical durian leaf classification in real orchard imagery benefits from architectural decisions that prioritize preservation of spatial detail and efficient multi-scale texture representation. The central meaning is that performance limitations in field symptom recognition are not necessarily resolved by increasing backbone depth or parameter count. Instead, the discriminative signal is better supported when early and mid-level features retain lesion morphology and when scale variability is addressed within the backbone using lightweight spatial operators. This interpretation is directly relevant to real-world deployment, where models must operate under constrained compute while remaining robust to heterogeneous acquisition conditions.

Several implications follow for model design and application. Firstly, for leaf image classification at moderate input sizes, aggressive late-stage downsampling can be counterproductive because it weakens fine-grained symptom evidence that is essential for separating visually similar classes. Secondly, multi-scale processing is most valuable when applied at stages where spatial structure remains explicit, since this is the point at which texture and boundary cues are encoded before being abstracted into higher-level semantics. Thirdly, efficiency should be evaluated beyond parameter count alone, because runtime is influenced by the placement of operations on large feature maps and by execution overhead from multi-branch modules. Together, these implications suggest that compact architectures tailored to symptom morphology can offer a more practical pathway for agricultural decision support than approaches that primarily rely on scaling model size.

Future research can extend these findings in three directions. The first direction is generalization and robustness. Evaluation across additional orchards, seasons, devices, and geographic regions is needed to quantify sensitivity to domain shift and to determine whether the architectural choices remain effective under broader operational variability. The second direction is statistical reliability and supervision. Multi-seed evaluation and significance testing across multiple data splits would strengthen the statistical basis of the reported improvements and clarify which performance differences are robust across random initializations. Additional supervision signals, such as attention consistency constraints or curriculum strategies that emphasize hard examples, may further reduce the ambiguities observed in failure cases. The third direction is deployment-oriented optimization. Operator-level profiling and kernel fusion can be explored to reduce multi-branch overhead, and lightweight post-training quantization or structured pruning can be investigated to further improve throughput without sacrificing accuracy. These directions provide a clear path toward more reliable and efficient durian leaf monitoring systems that can be adopted in routine orchard management.

REFERENCES

Balaha, H. M. (2026). A hybrid YOLOv8-EigenCAM model for accurate and interpretable plant disease classification. *Multimedia Tools and Applications*, *85*(2), 156. https://doi.org/10.1007/s11042-026-21382-8

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1704.04861

Jocher, G., Chaurasia, A., & Qui, J. (2023). *Ultralytics YOLOv8*. https://github.com/ultralytics/ultralytics

Jocher, G., & Qui, J. (2024). *Ultralytics YOLO11*. https://github.com/ultralytics/ultralytics

Joshi, K., Yadav, Y., Hooda, S., Nandal, R., Singh, B., Singh, K., Tuteja, N., Gill, R., & Gill, S. S. (2025). Classification of cotton leaf disease using YOLOv8 based k-fold cross validation deep learning method for precision agriculture. *Scientific Reports*, *15*(1), 35602. https://doi.org/10.1038/s41598-025-13147-4

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective Kernel Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 510–519. https://doi.org/10.1109/CVPR.2019.00060

Thang, N. D. (2025). Durian Leaf Disease Dataset. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/10611219

Thanh, T. N., Nguyen, L. X., Cap, T., & Le, T. (2025). A durian leaf image dataset of common diseases in Vietnam for agricultural diagnosis. *Data in Brief*, *61*, 111845. https://doi.org/10.1016/J.DIB.2025.111845

The Ministry of Finance of the Socialist Republic of Viet Nam, National Statistics Office. (2025, October 6). *Press release social-economic situation in the 3rd quarter and the 9 months of 2025*. https://www.nso.gov.vn/en/data-and-statistics/2025/12/press-release-social-economic-situation-in-the-3rd-quarter-and-the-9-months-of-2025/

Tian, Y., Ye, Q., & Doermann, D. (2025). *YOLO12: Attention-centric real-time object detectors*. https://github.com/sunsmarterjie/yolov12

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going deeper with Image Transformers. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 32–42). https://doi.org/10.1109/ICCV48922.2021.00010

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 6848–6856). https://doi.org/10.1109/CVPR.2018.00716

Zhao, J., Xu, L., Ma, Z., Li, J., Wang, X., Liu, Y., & Du, X. (2025). A review of plant leaf disease identification by deep learning algorithms. *Frontiers in Plant Science*, *16*, 1637241. https://doi.org/10.3389/FPLS.2025.1637241

Zhu, H., Wang, D., Wei, Y., Wang, P., & Su, M. (2025). YOLOV8-CMS: a high-accuracy deep learning model for automated citrus leaf disease classification and grading. *Plant Methods*, *21*(1), 88. https://doi.org/10.1186/s13007-025-01396-3