SYNTIA – Systems and Technologies in Artificial Intelligence Applications.
Volume 1 Issue 2 (June 2026). Page: 42-54.

Parallel CNN–Transformer Hybrid with Learnable Logit Fusion for Cocoa Pod Disease Classification

Dian Meldayani[1], Amriana[2*], Prakasit Poonwong[3]

[1] *Informatics Engineering Study Program, Faculty of Engineering, Universitas Tadulako, Palu, Indonesia.*
[2] *Department of Information Technology, Faculty of Engineering, Universitas Tadulako, Palu, Indonesia.*
[3] *Boromarajonani College of Nursing, Chainat, Nursing Faculty, Praboromrajchanok Institute, Thailand.*

**Corresponding author: amriana@untad.ac.id*

Abstract: Black pod rot and cocoa pod borer reduce cocoa yield, yet field diagnosis relies on scarce expertise. This study designs a parallel hybrid that couples a ResNet50 branch with a transformer branch on ResNet50 feature maps and fuses their class scores through a learnable weight. It was compared with ResNet50 and ViT-B/16 for three-class pod classification (healthy, black pod rot, pod borer) on 2,500 public images, split 70:15:15 before augmenting the training set only. Test accuracy was 97.60% (hybrid), 97.33% (ResNet50), and 97.07% (ViT-B/16; n = 375). The gaps equal one or two images, the 95% Wilson intervals overlap almost completely, and no paired exact test could reach significance, so no model can be declared superior. The hybrid made the fewest errors that labeled a diseased pod healthy (2, against 3 and 7) but needed 2.4 times the parameters of ResNet50. Grad-CAM on the convolutional branch gave broad, sometimes off-target maps, whereas attention rollout on the transformer branch focused on small spots; both were inspected qualitatively only. The hybrid is a competitive design with a clear computational cost; choosing between architectures needs repeated runs and field validation.

Keywords : cocoa pod disease; convolutional neural network; image classification; logit fusion; vision transformer.

Received: February 9, 2026 | Revised: March 24, 2026 | Accepted: April 21, 2026 | Published: June 30, 2026

1. Introduction

Pod diseases and pests limit the yield and quality of cocoa (*Theobroma cacao* L.). Black pod rot, caused by *Phytophthora* species, turns pods brown to black, and the cocoa pod borer (*Conopomorpha cramerella*) is a damaging pod pest that has been studied through beneficial insects and insectary plants [1]. Recognizing both problems early depends on trained observers and subjective visual judgment, which is why image-based deep learning has become the main route toward automated plant disease diagnosis [2].

Laboratory success does not guarantee field performance. A widely cited study reached 99.35% accuracy on a held-out set of 54,306 leaf images, but accuracy fell to just above 31% on images captured under conditions different from the training data [3]. For cocoa, where smallholders photograph pods under uncontrolled light and cluttered backgrounds, this gap makes both honest evaluation and interpretability important.

Research on cocoa pods has so far concentrated on convolutional neural networks (CNNs), as summarized in Table 1. Soh et al. compared five CNNs on 4,390 pod images [4]; Maylianti et al. reported 96% accuracy for both EfficientNet-B0 and ResNet-50 [5]; and Pradana et al. fine-tuned EfficientNet-B4 with seven augmentation types to reach 97.3% [6]. Ensembles of pretrained CNNs have reported near-ceiling accuracy on a balanced, augmented set of 6,000 pod images [7]. Work on other cocoa organs follows the same pattern: a CNN reached 99.1% on healthy versus vascular streak dieback leaves [8], and ResNet50

SYNTIA – Systems and Technologies in Artificial Intelligence Applications is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

outperformed VGG16 and a vision transformer at 94% on a public cocoa dataset [9]. Reported accuracies are not directly comparable because datasets, splits, and augmentation protocols differ.

Table 1. Representative studies on cocoa disease classification from images.

Study	Data	Model(s)	Acc. (%)
Soh et al. [4]	4,390 pod images	Five CNNs; best: custom CNN	91.79
Maylianti et al. [5]	4,390 pod images	EfficientNet-B0, ResNet-50	96
Pradana et al. [6]	Public pod images	EfficientNet-B4, fine-tuned	97.3
Anduyan et al. [7]	6,000 balanced, augmented pod images	Ensembles of six CNNs	up to 100
Jesse et al. [9]	Cocoa leaf images (CSSVD)	ResNet50, VGG16, ViT; best: ResNet50	94
Kumar et al. [8]	Healthy vs. VSD leaves	CNN	99.1
Kouassi et al. [10]	Pod images (swollen shoot symptoms)	CNN–ViT features + SVM/LightGBM	99.24
This study	2,500 pod images + field photographs	ResNet50, ViT-B/16, parallel hybrid	97.60

Class imbalance adds a second difficulty. Pod borer is scarce in public pod collections (103 of 4,390 images in one of them [5]), so the minority class is usually enlarged by augmentation [6]. That is safe only when the split is made first: otherwise near-identical copies of one photograph land in both the training and test sets and inflate accuracy.

Vision transformers (ViTs) model long-range relations between image patches [11], but they are usually trained on very large datasets, whereas CNNs bring a useful locality bias with little data. Hybrid CNN–ViT designs try to combine the two and are surveyed in [12] and [13]. A CNN–ViT hybrid has been applied to maize leaf diseases [14], and a hybrid trained from scratch on limited and imbalanced chest radiographs was reported to outperform both of its constituent models [15]. For cocoa pods, Kouassi et al. [10] used a CNN–transformer feature extractor followed by SVM and LightGBM classifiers and reported 99.24% accuracy on swollen-shoot symptoms.

Accuracy alone is also an incomplete basis for trust. Deep networks can respond to background cues instead of lesions, and explanation methods such as Grad-CAM and attention-based maps are therefore increasingly used to inspect what a vision model relies on [16]. Their faithfulness is itself debated, particularly for transformers [17], so such maps are best treated as diagnostic aids rather than proof. For smallholder cocoa growers, who commonly depend on extension officers for diagnosis and advice [18], a screening tool is useful only if it is cheap enough to deploy, reports its uncertainty, and fails in a predictable direction. These practical criteria, and not accuracy alone, frame the comparison made here.

Two gaps remain. First, among the studies reviewed here, none compares an end-to-end CNN–transformer hybrid with both of its constituent architectures on one fixed split while reporting sampling uncertainty. Second, with test sets of a few hundred images and accuracies near 97%, a difference between models is often one or two images, which a paired exact test can show to be indistinguishable from chance [19]; most reports do not address this.

This study therefore (i) designs a parallel hybrid whose CNN and transformer branches are fused by a learnable logit weight; (ii) compares it with ResNet50 and ViT-B/16 on one split together with parameters, operations, and latency; (iii) reports interval estimates and the limits of a paired test instead of ranking models by point accuracy; and (iv) inspects both branches with Grad-CAM and attention rollout on field photographs.

2. Method

The study follows the workflow in Fig. 1: a stratified split precedes augmentation, three models are trained on the same training images, and evaluation covers accuracy, uncertainty, error types, and computational cost. Field photographs enter only at the end, for qualitative interpretability checks.

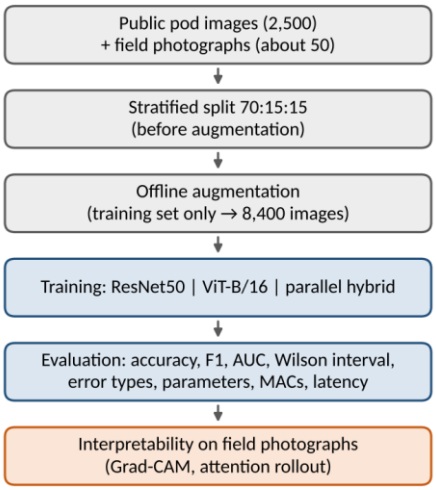


Figure 1. Research workflow.

2.1. Data

The 2,500 pod images come from a public Kaggle cocoa collection: 1,000 healthy, 1,000 black pod rot, and 500 pod borer images. Healthy pods have a normal husk color and a clean surface; black pod rot gives a brown husk; pod borer shows small dark specks on the surface. The data were divided into training, validation, and test sets (70:15:15, stratified) before any augmentation, so no augmented copy of a test or validation image can enter training (Table 2). Offline augmentation was applied to the training set only (random resized crop, horizontal flip, rotation, perspective, color jitter, Gaussian blur), three copies per image for healthy and black pod rot and seven for pod borer, yielding 2,800 images per class. Images were resized to 224×224 pixels and normalized with ImageNet statistics. No near-duplicate screening of the public images was performed.

Class definitions follow visible symptoms rather than laboratory confirmation. This is practical for a screening tool but it means that early or mixed infections, in which a pod shows both a faint lesion and borer specks, are forced into one label. Such ambiguous pods are probably the source of part of the residual error reported below, and a multi-label scheme would represent them more honestly than the single-label scheme used here.

Figure 2 shows one public and one field example per class. The field photographs contain cluttered backgrounds, bark, leaves, and uneven light that the public images largely lack, which is the domain shift that motivates the field checks below.

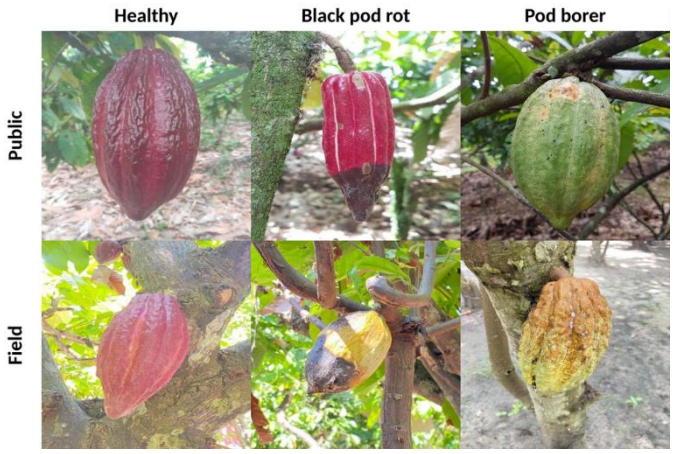


Figure 2. Example images of the three classes from the public collection (top) and from the field (bottom).

Table 2. Data partition (number of images).

Class	All	Train	Train (aug.)	Val.	Test
Healthy	1,000	700	2,800	150	150
Black pod rot	1,000	700	2,800	150	150
Pod borer	500	350	2,800	75	75
Total	**2,500**	**1,750**	**8,400**	**375**	**375**

In addition, about 50 photographs of pods were taken in cocoa plantations in Central Sulawesi with a smartphone camera under natural light, and their labels were confirmed by a local agricultural extension officer. These field images were never used for training, validation, or model selection; they serve only for the qualitative checks in Section 3.4.

2.2. Models

The first baseline is an ImageNet-pretrained ResNet50 [20] whose final layer is replaced by a three-class head (2048→512→3, dropout 0.3). The second is a ViT-B/16 [11]. The hybrid has two parallel branches (Fig. 3). The CNN branch is the ResNet50 above. The transformer branch passes the image through a separate pretrained ResNet50 up to its third stage, which produces a 14×14×1024 feature map; a 1×1 convolution projects it to 768 channels, the 196 positions become tokens, a class token and learnable positional embeddings are added, and three transformer encoder blocks (12 heads, MLP ratio 4, dropout 0.1, stochastic depth up to 0.05) process the sequence. The class token feeds a head (768→256→3). The two sets of logits are combined by

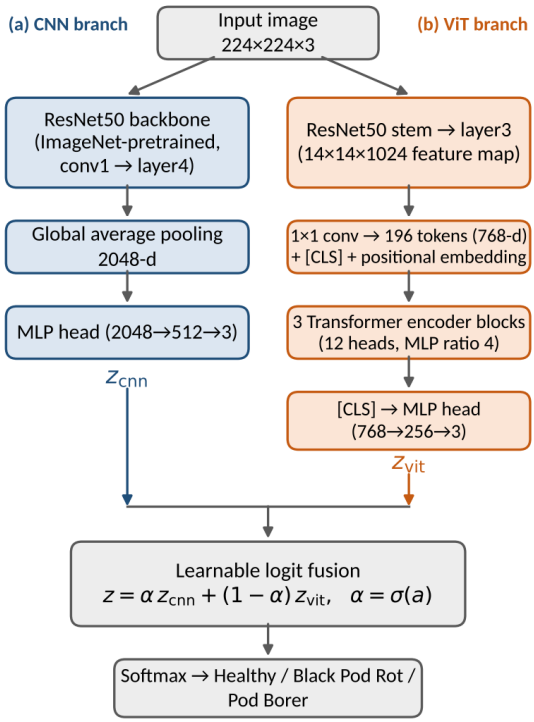


Figure 3. Architecture of the parallel hybrid. The CNN branch (a) and the transformer branch (b) produce class scores that are fused with a single learnable weight.

$$\mathbf{z} = \alpha\, \mathbf{z}_{\mathrm{cnn}} + (1-\alpha)\, \mathbf{z}_{\mathrm{vit}} \tag{1}$$

where $\mathbf{z}_{\mathrm{cnn}}$ and $\mathbf{z}_{\mathrm{vit}}$ are the logits of the two branches and α is a scalar weight obtained from a learnable parameter a through

$$\alpha = \sigma(a) = \frac{1}{1+e^{-a}} \tag{2}$$

The parameter a is initialized to 1.2 (α ≈ 0.77) and trained jointly with the network; α stays in (0, 1) by construction.

2.3. Training protocol

The hybrid was trained for at most 30 epochs in two stages with the AdamW optimizer [21] (weight decay 0.05), a cosine learning-rate schedule, cross-entropy loss, and gradient clipping. In stage A (8 epochs, learning rate 1×10^{-4}) both backbones were frozen and only the heads, the projection, the transformer blocks, the class token, the positional embeddings, and a were trained. In stage B (22 epochs) the last CNN stage and the last stage of the ResNet50 inside the transformer branch were unfrozen at 1×10^{-5}, the other trainable parts used 5×10^{-5}, and training stopped early if validation accuracy did not improve for 10 epochs. The checkpoint with the best validation accuracy was kept. Each model was trained once. Experiments ran in PyTorch on a laptop with an NVIDIA RTX 4050 GPU, an Intel Core i7 processor, and 12 GB of memory.

2.4. Evaluation and statistics

Models were evaluated on the 375 test images with accuracy, macro-averaged precision, recall, and F1-score, one-vs-rest ROC-AUC, test loss, and confusion matrices. Uncertainty in accuracy is expressed by the 95% Wilson score interval [22],

$$\mathrm{CI}_{95\%} = \frac{\hat{p}+\frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}}{1+\frac{z^2}{n}} \tag{3}$$

where $\hat{p}$ is the observed accuracy, n = 375 is the number of test images, and z = 1.96. Two classifiers evaluated on the same images are properly compared with an exact McNemar test on the counts b and c of images that only one of them classifies correctly [19]; the two-sided p-value is

$$p = \min\left\{1,\ 2\sum_{i=0}^{\min(b,c)}\binom{b+c}{i}\left(\frac{1}{2}\right)^{b+c}\right\} \tag{4}$$

The Wilson interval is preferred to the simple normal interval because accuracies near 100% on a few hundred images make the normal interval too narrow and able to exceed 100% [22]. Intervals are reported for accuracy and for the miss rate of diseased pods, the two quantities on which conclusions are drawn. No correction for multiple comparisons was applied, because no hypothesis test could be completed and all intervals are descriptive.

Per-image predictions were not stored, so the test could not be run; Section 3.2 instead bounds what Eq. (4) could show from the accuracy gap. Complexity was measured as trainable parameters, multiply–accumulate operations (MACs) from the thop profiler, and mean latency per image (batch size 1; 100 timed forward passes after 10 warm-up passes.

2.5. Interpretability

Grad-CAM [23] weights the activations of a convolutional layer of the CNN branch by the spatially averaged gradients of the score of the predicted class, which here is taken from the fused logits, and keeps the positive part of the weighted sum. For the transformer branch, attention rollout [24] averages each block's attention matrix over heads, adds the identity to account for the residual connection, renormalizes rows, and multiplies the three block matrices:

$$\tilde{A}_l = \mathrm{norm}(\bar{A}_l + I), \qquad R = \tilde{A}_3\tilde{A}_2\tilde{A}_1 \tag{5}$$

where $\bar{A}_l$ is the head-averaged attention of block l, I is the identity matrix, and norm(·) divides each row by its sum. The row of R belonging to the class token, with the class-token column removed, gives one relevance value for each of the 196 positions; it is reshaped to 14×14 and min–max scaled for display. Both maps were upsampled to the input size and overlaid on the photograph.

2.6. Error-cost view and metric definitions

Besides overall accuracy, the study treats the error types separately. A pod that is diseased but classified as healthy (a false negative for disease) delays action and may spread infection, whereas a healthy pod flagged as diseased mainly costs an unnecessary inspection. For this reason the results report, for each model, the number of diseased test pods predicted as healthy together with its Wilson interval. Per-class precision, recall, and F1-score follow the usual definitions, and the macro F1-score is their unweighted mean,

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K}\frac{2P_kR_k}{P_k+R_k}, \qquad K = 3 \tag{6}$$

where P_k and R_k are the precision and recall of class k. Macro averaging gives the 75 pod borer images the same weight as the 150 images of each other class, which suits the imbalanced test set better than accuracy alone.

2.7. Reproducibility and ethics

The split was drawn once with stratification and then fixed for all models, so every model saw identical training, validation, and test images. Checkpoint selection used the validation set only. All experiments used a fixed random seed of 42 across NumPy, PyTorch, and CUDA operations to ensure reproducibility. Models were implemented in PyTorch 2.1 with torchvision 0.16, consistent with the training hardware described in Section 2.3 (NVIDIA RTX 4050 GPU). The code and trained weights are available from the corresponding author upon reasonable request. Verbal permission to photograph pods in the plantations and

to use the images for this study was obtained from the plantation owners in Central Sulawesi before the field photographs described in Section 2.1 were taken.

2.8. Decision-support prototype

A prototype wraps the trained hybrid. After classification it shows the class probabilities together with a short description of the condition, its likely cause, and recommended handling for the predicted class. The text was compiled from an explanation given by a local agricultural extension officer and from a community-extension article [18]; it is a fixed lookup, not a learned component.

3. Results and Discussion

3.1. Training behavior

The hybrid reached its best validation accuracy of 97.87% (367 of 375 images) before early stopping ended training. The fusion weight barely moved from its initial value ($\alpha \approx 0.769$ at initialization, 0.759 after training), so about three quarters of the weight stayed on the CNN branch. Because α changed so little, it should not be read as evidence that the CNN branch carries more useful information than the transformer branch; it mainly reflects the initialization and the short second stage.

3.2. Classification performance

Table 3 shows that the three models perform at the same level. The hybrid has the highest accuracy, macro F1-score, and macro AUC, but ResNet50 has the highest macro precision and clearly the lowest test loss (0.1283 against 0.1911 for the hybrid). The accuracy gaps are 0.27 percentage points (one image) between the hybrid and ResNet50 and 0.53 points (two images) between the hybrid and ViT-B/16.

Table 3. Test-set performance of the three models (n = 375).

Model	Acc. (%)	Prec.	Rec.	F1	AUC	Loss
Hybrid	97.60	0.9736	0.9778	0.9756	0.9983	0.1911
ResNet50	97.33	0.9757	0.9733	0.9744	0.9974	0.1283
ViT-B/16	97.07	0.9736	0.9733	0.9733	0.9972	0.1555

Prec., Rec., F1, AUC: macro averages; Loss: cross-entropy on the test set.

The 95% Wilson intervals (Eq. 3) are about 3 points wide, roughly ten times the gaps between the models. A paired exact test cannot rescue the ranking: with a net difference of one image the smallest possible two-sided p-value from Eq. (4) is 1.0, and with a net difference of two images it is 0.5. Even if one model were right on every image where the other erred ($c = 0$), at least six such images would be needed for $p < 0.05$. The data therefore support the statement that the three models perform equally well on this test set, not that the hybrid is best. This differs from the conclusion that the hybrid is the best of the three that would follow from the point estimates alone, and it is consistent with the narrow spread among CNN-only cocoa studies (Table 1), where the ResNet50 baseline here (97.33%) already sits at the upper end.

Table 4. Per-class recall and F1-score on the test set.

Class (n)	Rec. Hyb.	Rec. Res.	Rec. ViT	F1 Hyb.	F1 Res.	F1 ViT
Black pod rot (150)	0.973	0.980	0.953	0.973	0.970	0.966
Healthy (150)	0.973	0.967	0.980	0.980	0.973	0.967
Pod borer (75)	0.987	0.973	0.987	0.974	0.980	0.987

Rec.: recall; Hyb.: hybrid; Res.: ResNet50; ViT: ViT-B/16.

Per-class results (Table 4) also fail to single out one model. The hybrid has the best F1-score on healthy pods, ResNet50 on none of the classes by a clear margin, and ViT-B/16 reaches the highest F1-score of any cell (0.987) on pod borer, the smallest class. With only 75 pod borer images, one image moves recall by 1.3 percentage points, so these differences carry little information. One-vs-rest AUC exceeded 0.993 for every class and model, which shows that the classes are well separated in score space by all three networks.

How large would a gap have to be to count as evidence? For the exact test in Eq. (4), the smallest net difference in correct images that reaches $p < 0.05$ for a given number of discordant images, that is, images that exactly one of the two models classifies correctly. Fewer than six discordant images can never reach significance. With 10 to 12 discordant images the net difference must be 8 images (2.1 percentage points at n = 375), and with 30 it must be 12 images (3.2 points). The observed gaps of one and two images are far below these thresholds.

3.3. Error structure

The confusion matrices in Fig. 4 are summarized by error type in Table 5, which matters more agronomically than the headline accuracy. Calling a diseased pod healthy delays treatment. Such errors numbered 2 for the hybrid, 3 for ResNet50, and 7 for ViT-B/16 (of 225 diseased test pods), with ViT-B/16 mostly confusing black pod rot with healthy pods. These counts favor the hybrid, but they are small and come from one training run, so they are a hypothesis for repeated runs rather than a finding.

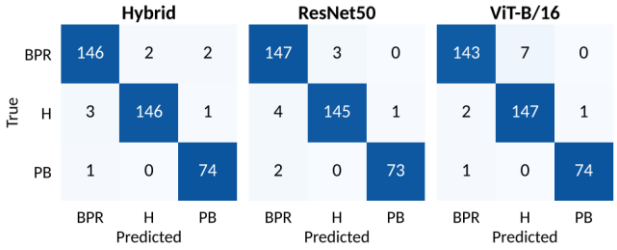


Figure 4. Confusion matrices on the test set (BPR: black pod rot; H: healthy; PB: pod borer).

Expressed as rates, the miss rates for diseased pods are 0.9% (2/225) for the hybrid, 1.3% (3/225) for ResNet50, and 3.1% (7/225) for ViT-B/16. Their 95% Wilson intervals are 0.2–3.2%, 0.5–3.9%, and 1.5–6.3%. The hybrid and ResNet50 intervals overlap almost completely, and even the interval for ViT-B/16 touches the upper ends of the other two. The apparent advantage of the hybrid on this criterion is therefore again within sampling noise, although the direction of the difference is consistent across the three comparisons. A further pattern concerns pod borer: no model labeled a borer-damaged pod as healthy, and at most two of 75 were assigned to black pod rot, so confusion with the healthy class arises almost entirely from black pod rot, whose early lesions can resemble a healthy husk.

Table 5. Misclassified test images by error type (true → predicted).

Error type	Hybrid	ResNet50	ViT-B/16
Black pod rot → healthy	2	3	7
Black pod rot → pod borer	2	0	0
Healthy → black pod rot	3	4	2
Healthy → pod borer	1	1	1
Pod borer → black pod rot	1	2	1
Pod borer → healthy	0	0	0
Total errors (of 375)	**9**	**10**	**11**

3.4. Complexity, field images, and interpretability

Table 6 shows what the extra branch costs. Relative to ResNet50, the hybrid has 2.4 times the parameters, 2.9 times the MACs, and 2.2 times the latency, for a gain of one test image. MACs and latency disagree in rank (the hybrid has slightly more MACs than ViT-B/16 but 28% lower latency), which suggests that the profiler missed operations such as attention products and that memory behavior matters, so the MACs column is approximate and a hardware-specific benchmark is needed. Absolute latencies (hundreds of milliseconds) are also higher than expected for a modern GPU, which is why the measurement device is flagged above. On present evidence ResNet50 is the more economical choice for a phone-side or low-cost screening tool.

Table 6. Computational cost of the models.

Model	Params (M)	MACs (G)	Latency (ms)
ResNet50	23.51	4.13	245.93
ViT-B/16	85.80	11.29	763.88
Hybrid	55.51	11.79	552.72

MACs from the thop profiler; latency per image at batch size 1.

Field photographs expose the lab-to-field gap discussed in the Introduction [3]. On ten field photographs inspected qualitatively, the hybrid classified all ten correctly, and ResNet50 and ViT-B/16 eight each. Ten images cannot rank models (the lower limit of the 95% Wilson interval for 10 of 10 is 72%), so a quantitative evaluation on the roughly 50 field images, with models trained only on public data, is the most important missing experiment. The per-class probability plots for these photographs were softened with a fixed temperature (T = 3) that was not fitted on held-out data, so they illustrate relative confidence only and are not calibrated probabilities.

Fig. 5 shows both interpretability maps on four field photographs. Grad-CAM from the CNN branch is broad: in the first row it peaks on the left edge of the front pod and on the background, in the second its strongest response lies on the neighboring pod rather than on the decayed pod itself, in the third it forms a wide band across the pod body, and in the fourth it peaks at the lower border below the pods. Attention rollout from the transformer branch is sparser, marking a few small spots, although some fall on bark or leaves. The two maps therefore differ in granularity. They were not compared with lesion annotations, and attention rollout approximates information flow and is not a faithful explanation by construction [24], [17], [16]. The statement that the CNN branch attends to broad texture and the transformer branch to symptom spots should thus be read as a description of four examples. That the strongest Grad-CAM response lies outside the pod of interest in two of the four (second and fourth rows) also suggests that background cues can influence predictions, which argues for cropping or background suppression before classification.

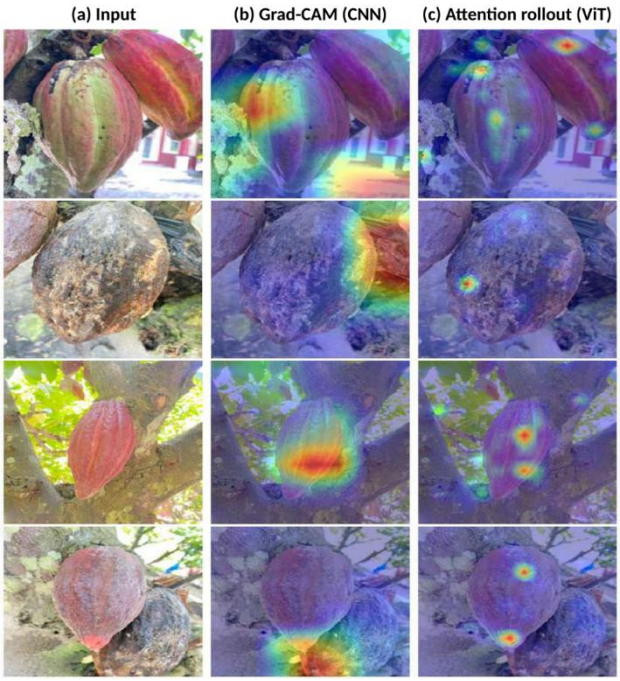


Figure 5. Four field photographs (a) with Grad-CAM of the CNN branch (b) and attention rollout of the transformer branch (c) for the hybrid model.

3.5. Decision-support use

The prototype maps the predicted class and its probabilities to a short text on cause and handling (sanitation, pruning, and biological agents or fungicide), compiled from the extension officer's explanation and a community-extension article [18]. This content has not been validated by agronomists. Given the

error structure above, a deployed tool should show the two most probable classes, send low-confidence images to a human, and make clear that a healthy prediction is not a guarantee, because calling a diseased pod healthy was the costly error in every model. Pod borer management through beneficial insects and insectary plants is an active research topic [1], so recommendations tied to the detected pest can later be linked to such findings.

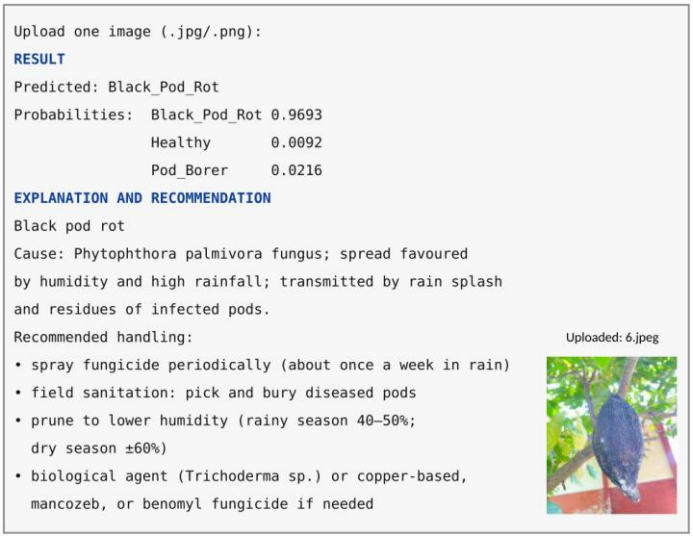


Figure 6. Output of the recommendation prototype for one field photograph

Fig. 6 shows the prototype output for a black pod rot photograph, classified with probability 0.9693 against 0.0092 for healthy and 0.0216 for pod borer. The prototype is a notebook interface in which the user uploads one image; it prints the class probabilities and a fixed text block for the predicted class, so the recommendation is a lookup in a small knowledge base keyed by the class and not a decision computed from the image. Three cautions follow from the screenshot. First, the percentages in the pruning advice (rainy season 40–50%, dry season about 60%) are not defined in the source, and it is unclear whether they refer to shade cover or to humidity. The community-extension article [18], the source compiled for the prototype's advisory text, discusses pruning in terms of canopy shade, so shade cover is the more likely reading. Second, the listed fungicides, copper-based products, mancozeb, and benomyl, should be checked against the current Indonesian pesticide registration before the text is shown to growers, because some active ingredients have been withdrawn in several countries. Third, the probability shown is the network output, which for an image that is not a cocoa pod would still be assigned to one of the three classes; a rejection option for unfamiliar images, as suggested below, is needed before the tool is offered outside a research setting.

3.6. Position relative to earlier studies

The accuracy of the hybrid (97.60%) is close to that of EfficientNet-B4 (97.3%) [6] and above those reported in [5] (96%) and [4] (91.79%), and the ResNet50 baseline here (97.33%) is just as high. Such comparisons are only indicative: class counts differ (only 103 pod borer images in the data of [5]), splits and augmentation differ, and the near-perfect figures of [7] and [10] come from other datasets. Hybrid gains reported for chest radiographs trained without pretraining [15] concern a different modality and training regime, so the present result neither confirms nor contradicts them. What it does show is that on a 375-image pod test set a hybrid advantage of the size usually reported in this literature (about one percentage point) would span a handful of images and could not be separated from sampling noise without repeated runs.

3.7. Threats to validity and future work

Several limitations qualify the results. Each model was trained once, so seed-to-seed variance is unknown. The test set has 375 images, only 75 of them pod borer, and public collections may contain near-duplicates across splits that no one has screened for. The baselines may have received less tuning effort than the hybrid, which would bias the comparison in its favor. The latency device was not recorded and MACs came from a profiler that may under-count operations. Field images were not scored quantitatively, and interpretability was assessed visually on a handful of photographs.

A further limitation is domain shift. The public images were photographed for dataset collection, with the pod usually centered and the background simplified, whereas the field photographs show clutter and variable light. Good test accuracy on the public split therefore says little about the accuracy a farmer would obtain, and the Grad-CAM peaks outside the pod in Fig. 5 point in the same direction. Finally, the three-class scheme is narrow: other cocoa conditions, such as vascular streak dieback or mirid damage, are not represented, so a pod with an unseen condition will still be assigned to one of the three known classes with apparent confidence.

Future work should (i) repeat training over several seeds or stratified folds and report mean and spread; (ii) store per-image predictions so that paired McNemar tests become possible; (iii) add ablations with a fixed α and without the transformer branch to learn whether the second branch contributes at all; (iv) evaluate the roughly 50 field photographs quantitatively and test background suppression; and (v) try lighter transformer backbones, since the present hybrid's cost is its main practical drawback.

4. Conclusion

The parallel CNN–transformer hybrid with learnable logit fusion classified cocoa pods into three classes with 97.60% test accuracy, against 97.33% for ResNet50 and 97.07% for ViT-B/16. Because the gaps equal one or two of 375 images, the intervals overlap almost entirely, and a paired exact test could not reach significance, the study supports competitiveness of the hybrid, not superiority. Its one possibly useful property, fewer diseased pods called healthy, rests on counts of 2, 3, and 7 from a single run. The hybrid costs 2.4 times the parameters of ResNet50 and 2.2 times its latency, so ResNet50 remains the practical default for resource-limited screening until repeated runs and a quantitative field evaluation show a real gain. For practice, systems of this kind should report interval estimates, expose class probabilities, and be validated on field photographs before farmers rely on them.

Acknowledgments

The authors thank Arman Herianto, S.P., agricultural extension officer, for confirming field labels and explaining cocoa pod disease management.

AI Disclosure Statement

The authors used Generative AI and AI-Assisted Technologies throughout the preparation of this manuscript: to draft and revise portions of the text, to write and run the scripts behind the experiments and statistical analyses, to generate the figures included in the revision, and to format the document from the authors' own research data and source code. The authors verified the scripts and their outputs, reviewed the resulting text, cross-checked every reference against its original source, and accept full responsibility for the content of this publication.

Author Contributions Statement

Dian Meldayani: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Data Curation, Visualization, Writing – Original Draft.
Amriana: Supervision, Validation, Writing – Review and Editing.
Prakasit Poonwong: Validation, Writing – Review and Editing.

References

[1] V. S. Dewi et al., "The presence of beneficial insects and damage intensity of cocoa pod borer (*Conopomorpha cramerella* Snellen) in plantations with and without insectary plants," *J. Trop. Plant Pests Dis.*, vol. 25, no. 2, pp. 179–189, 2025, doi: 10.23960/j.hptt.225179-189.

[2] J. Liu and X. Wang, "Plant diseases and pests detection based on deep learning: A review," *Plant Methods*, vol. 17, no. 1, Art. no. 22, Feb. 2021, doi: 10.1186/s13007-021-00722-9.

[3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Front. Plant Sci.*, vol. 7, Art. no. 1419, Sep. 2016, doi: 10.3389/fpls.2016.01419.

[4] K. S. Soh, E. G. Moung, K. J. J. Danker, J. A. Dargham, and A. Farzamnia, "Cocoa diseases classification using deep learning algorithm," *ITM Web Conf.*, vol. 63, Art. no. 01014, 2024, doi: 10.1051/itmconf/20246301014.

[5] N. P. Maylianti, I. G. N. L. Wijayakusuma, and I. P. C. A. Wiguna, "Comparison of EfficientNet-B0 and ResNet-50 for detecting diseases in cocoa fruit," *J. Appl. Informatics Comput.*, vol. 9, no. 1, 2025, doi: 10.30871/jaic.v9i1.8868.

[6] A. G. Pradana, D. R. I. M. Setiadi, and A. R. Muslikh, "Fine tuning model convolutional neural network EfficientNet-B4 dengan augmentasi data untuk klasifikasi penyakit kakao," *J. Inf. Syst. Appl. Dev.*, vol. 2, no. 1, pp. 1–11, Mar. 2024, doi: 10.26905/jisad.v2i1.11899.

[7] D. Anduyan, N. Cabillo, N. Gultiano, and M. P. Pacot, "Enhancing cocoa pod disease classification via transfer learning and ensemble methods: Toward robust predictive modeling," arXiv:2504.12992, Apr. 2025.

[8] S. Kumar, A. K. Sharma, S. Srivastava, V. Tiwari, P. K. Patidar, and S. Rai, "Automated cocoa disease detection using convolutional neural networks: A case study of VSD and other pathogens," in *Proc. Int. Conf. Recent Adv. Modernization Sustain. Intell. Technol. Appl. (RAMSITA-2025)*, Adv. Intell. Syst. Res., vol. 192, 2025, doi: 10.2991/978-94-6463-716-8_29.

[9] A. Jesse, N. Y.-R. Douha, and P. Lenka, "Image classification for CSSVD detection in cacao plants," arXiv:2405.04535, 2024.

[10] K. S. Kouassi, M. Diarra, K. H. Edi, and K. B. Jean-Claude, "Detection of cocoa pod diseases using a hybrid feature extractor combining CNN and vision transformer with dual classifier," *Edelweiss Appl. Sci. Technol.*, vol. 9, no. 1, pp. 668–681, 2025, doi: 10.55214/25768484.v9i1.4209.

[11] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, arXiv:2010.11929.

[12] H. Yunusa, S. Qin, A. H. A. Chukkol, A. A. Yusuf, I. Bello, and A. Lawan, "Exploring the synergies of hybrid CNNs and ViTs architectures for computer vision: A survey," *Eng. Appl. Artif. Intell.*, Art. no. 110057, 2025, doi: 10.1016/j.engappai.2025.110057.

[13] R. Azad et al., "Advances in medical image analysis with vision Transformers: A comprehensive review," *Med. Image Anal.*, vol. 91, Art. no. 103000, Jan. 2024, doi: 10.1016/j.media.2023.103000.

[14] G. Shandilya, S. Gupta, H. G. Mohamed, S. Bharany, A. U. Rehman, and S. Hussen, "Enhanced maize leaf disease detection and classification using an integrated CNN-ViT model," *Food Sci. Nutr.*, vol. 13, no. 7, Art. no. e70513, Jul. 2025, doi: 10.1002/fsn3.70513.

[15] P. S. Basnet and R. Chitrakar, "CNN-ViT hybrid for pneumonia detection: Theory and empiric on limited data without pretraining," arXiv:2509.08586, Sep. 2025.

[16] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, "A comprehensive review of explainable artificial intelligence (XAI) in computer vision," *Sensors*, vol. 25, no. 13, Art. no. 4166, Jul. 2025, doi: 10.3390/s25134166.

[17] S. Stassin, V. Corduant, S. A. Mahmoudi, and X. Siebert, "Explainability and evaluation of vision transformers: An in-depth experimental study," *Electronics*, vol. 13, no. 1, Art. no. 175, Jan. 2024, doi: 10.3390/electronics13010175.

[18] E. F. Muniroh, G. Safitri, S. D. Fadilah, and S. Sa'diyah, "Pemberdayaan kelompok tani melalui penyuluhan budidaya tanaman kakao dan pengendalian hama penyakit kakao," *Lembaran Masyarakat*, vol. 6, no. 1, Jun. 2020, doi: 10.32678/lbrmasy.v6i1.3711.

[19] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, arXiv:1711.05101.

[22] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Stat. Sci.*, vol. 16, no. 2, pp. 101–133, May 2001, doi: 10.1214/ss/1009213286.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[24] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2020, pp. 4190–4197, doi: 10.18653/v1/2020.acl-main.385.