

Article

Saturated In-Domain, Separable Out-of-Domain: A Hierarchical Multi-Scale Lesion-Attention Network and a Zero-Shot Cross-Corpus Protocol for Multi-Crop Leaf Disease Classification

Songul Karakus , Mehmet Burukanli  and Davut Ari 

Department of Computer Engineering, Faculty of Engineering and Architecture, Bitlis Eren University, Bitlis 13000, Türkiye; mburukanli@beu.edu.tr (M.B.); dari@beu.edu.tr (D.A.)

* Correspondence: skarakus@beu.edu.tr

Abstract

Leaf disease classifiers are ranked by held-out accuracy on their training corpus, where that number now exceeds 98% for many modern backbones. We ask what such a measurement can still resolve. On a four-crop, 21-class corpus of 7179 de-duplicated images, we train 26 architectures under one protocol with three seeds and shared partitions, and we execute the entire benchmark twice. The 26 architectures span 1.68 percentage points (pp) of internal macro-F1. The same configuration differs by 0.34 pp on average and by up to 1.06 pp between the two executions, and the internal ordering only partly replicates (Spearman's $\rho = 0.69$, with VGG-19 moving from 17th to 1st). Applied zero-shot to an independently collected mango corpus, the same checkpoints spread over 47 pp of accuracy and their ordering does replicate ($\rho = 0.92$). We propose HMLA-Net, which adds multi-scale fusion, gated multi-head lesion-aware pooling, an auxiliary crop head and MixStyle to a CAFormer-S18 backbone and averages the posterior over the Klein four-group of flips at inference. In the reference execution, it attains the highest internal macro-F1 (0.9845) and the highest zero-shot accuracy (0.699) of the 26 architectures, with the smallest degradation (28.9 pp) and the lowest rate of predictions leaking into labels absent from the target corpus. The second execution reproduces its zero-shot rank but not its internal one. Five backbones retrained with its complete recipe show that its margin over its own backbone lies within replicate variation, and in a 16-removal ablation measured on both partitions, only the large effect of ImageNet pretraining remains consistent across the two executions. A cross-corpus duplicate audit finds no exact duplicates and no evidence of substantial overlap, and one class, mango sooty mould, collapses to near-zero recall in 19 of 26 architectures. The contribution is the protocol: replicate the benchmark, evaluate out of domain, and report results per class.

Keywords: plant disease classification; cross-dataset generalisation; zero-shot transfer; benchmark saturation; replicability; symmetry-invariant inference; multiple-instance pooling; domain randomisation



Academic Editor: Jie Yang

Received: 10 August 2026

Revised: 7 September 2026

Accepted: 15 September 2026

Published: 17 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Plant pests and diseases remain one of the largest recurring losses in global food production. The Food and Agriculture Organization estimates that up to 40% of global crop production is lost annually to plant pests and diseases, with associated trade losses exceeding USD 220 billion each year [1]. Because plants supply roughly 80% of the human

diet, these losses translate directly into food-security pressure [2], and plant pathogens continue to spread into new regions as the climate changes [1,3]. Early identification of foliar symptoms is therefore a practical lever on yield, and one well matched to image-based automation.

Deep learning has been applied to this problem intensively, and on curated benchmarks, the task appears solved, since accuracies above 99% are routine. Our own measurements reproduce that impression and then undermine it. Training 26 architectures on a four-crop, 21-class corpus under a single protocol with shared partitions, we find every one of them between 0.968 and 0.985 macro-F1, a total spread of 1.68 percentage points across designs that differ by two orders of magnitude in parameter count and by fundamental inductive bias. To learn how much of that spread is signal, we did what is rarely done and executed the whole benchmark a second time, with the same code, hyper-parameters, partitions and hardware. The same configuration on the same seed and split differs between the two executions by 0.34 percentage points on average and by 1.06 at most (0.19 points for the three-seed means that the tables report), which is a fifth of the entire spread and larger than the gap between the first and the second model in either execution. The two orderings agree only partly (Spearman's $\rho = 0.69$). VGG-19 moves from 17th to 1st, the proposed model from 1st to 8th and ConvNeXt-S from 6th to 17th. The seed-level significance tests that this literature conventionally reports cannot add information here, because with three paired blocks, the smallest attainable two-sided Wilcoxon p value is 0.25 and the Nemenyi critical difference spans 92% of the rank axis. We say so in advance, and we keep that statement separate from the empirical one, since an underpowered test is not evidence of equivalence.

A benchmark whose spread is comparable to its own replicate-to-replicate variation cannot order its leading methods, and a paper reporting that a new architecture leads such a benchmark by tenths of a point may be reporting a difference that would not survive re-execution. This is not an argument that architecture does not matter. It is an argument that the wrong quantity is being measured. When we take the same 26 checkpoints and apply them, without any adaptation, to a mango corpus assembled by a different group with a different camera, the same models spread over 47.3 percentage points of accuracy and 36.9 of macro-F1, and this ordering does replicate between the two executions ($\rho = 0.92$ for accuracy), even though the individual numbers are noisier in absolute terms. The external corpus discriminates among the same configurations far more strongly than the internal partition does.

Three concrete gaps follow, and they define this work. The first concerns evaluation. Genuine zero-shot cross-corpus evaluation is rare. Of the twenty-five method papers reviewed in Section 2, none performs it on an independently collected corpus, while several use the term “cross-dataset” for protocols that retrain or fine-tune on the target corpus. Where transfer has been measured, the reported collapses are severe, with macro-F1 falling from 0.987 to 0.012 across tomato corpora [4] and accuracy falling by 14 to 24 points across four backbones moving from web-sourced images to PlantDoc [5]. The second concerns scope. Reviews note that existing models “are only capable of detecting particular crops” [6], and the multi-class studies that exist rely overwhelmingly on PlantVillage, whose fixed laboratory acquisition is itself a documented source of bias [7,8]. The third concerns what ablation studies establish. Component ablations are almost invariably measured in-domain, on precisely the partition we show to be saturated, and almost never replicated. We show below that at three seeds, a component ablation attributes nothing on either partition except the effect of pretraining, and that its out-of-domain ordering does not survive re-execution.

We address all three. Our contributions are as follows.

- We benchmark 26 architectures on a four-crop, 21-class corpus, de-duplicated by content hash, with three seeds and a partition shared by every model within a seed, and we execute the entire benchmark twice. The 78 paired repeat runs give a distribution of replicate-to-replicate differences, not a single number, and show that the in-domain ordering is itself only partly reproducible.
- We propose HMLA-Net, a hierarchical multi-scale lesion-attention network that adds four components to a CAFormer-S18 backbone. These are multi-scale fusion of three backbone stages onto a stride-16 grid with learned scale weights, gated multi-head lesion-aware attentive pooling in the multiple-instance sense, an auxiliary crop head whose combination coefficient is calibrated on validation over a grid containing zero, and MixStyle randomisation of feature statistics in the two convolutional stages. The prediction is made exactly invariant under the Klein four-group of flips that the imaging geometry admits, by averaging the posterior over the group orbit at inference. Four-view flip averaging is a known technique, and our contribution is its formal statement and its measurement under distribution shift. In the reference execution, HMLA-Net attains the highest internal macro-F1 and the highest zero-shot accuracy of the 26 architectures, with the smallest internal-to-external degradation, at 23.9 M parameters. The second execution reproduces the zero-shot rank and the degradation, but not the internal rank, which we therefore report as a fact about one execution.
- We evaluate all checkpoints on MLD24 with the 21-way head unchanged, no fine-tuning and no target-domain statistic of any kind, and we report three scoring protocols together with the fraction of predictions that leak into the 15 labels MLD24 does not contain. To separate architecture from recipe, five of the strongest backbones are retrained with the complete HMLA-Net recipe. Under identical recipes, the margin of HMLA-Net over its own backbone is inside replicate variation, and we make no architectural claim beyond that.
- We measure a 16-removal ablation on both partitions and in both executions, and show that only ImageNet pretraining produces an effect larger than replicate variation and stable across executions, that the out-of-domain ordering of the other components does not replicate ($\rho = 0.34$ between executions), and that the in-domain ablation yields a negative estimated contribution for the majority of components in both.
- We audit the two corpora for exact and near-duplicate images, finding no exact duplicates and no evidence of substantial overlap, and we report a corpus-level finding. Mango sooty mould is classified perfectly in-domain and collapses to near-zero recall under transfer in 19 of the 26 architectures. We present label definition, disease stage and acquisition as non-exclusive hypotheses for this collapse, with quantitative image descriptors and a visual comparison, rather than as an established fact.

2. Related Work

Transfer learning from ImageNet-pretrained convolutional backbones remains the default approach to leaf disease classification, and on curated single-crop corpora, it is close to saturated. Ahmmed et al. [9] fine-tuned five pretrained CNNs on an eight-class mango corpus, of which DenseNet201 reached the highest accuracy, 99.33%, ahead of ResNet152V2, SeResNet152, InceptionV3 and Xception. Gopalan et al. [10] reported 98.34% with ResNet152 on a four-class maize set, and Alhammad et al. [11] 98% with a frozen-backbone VGG16 on the three-class PlantVillage potato subset. For the remaining crops in our corpus, the literature is thinner. Cashew disease recognition has been addressed, among other approaches, through generative augmentation and

TinyML-oriented model compression [12], and jackfruit by only a few studies, among them Tuteja et al. [13], whose CNN–attention–LSTM model reaches 99% over three classes. Two properties of this body of work limit how far its numbers can be compared. Reported accuracies are frequently measured on evaluation sets containing augmented or synthetic copies of training images. Augmentation precedes the split in [9,11]; the source corpus was already augmented by its creator in [13]; and in [12], GAN-generated images entered the evaluation data: the same architecture (MobileNetV2) scored 97.7% on the GAN-augmented set but 96% when trained and tested on the original images alone (see Table 1). Test partitions are also often small enough that a single misclassification moves the headline figure appreciably. Hossain et al. [14] evaluate on a test partition of 400 images (10% of the 4000-image corpus), so 99.75% corresponds to one error.

Vision transformers were introduced to this domain to capture the long-range dependencies that convolutions model only indirectly. Hossain et al. [14] showed that an optimised DeiT reaches 99.75% on MangoLeafBD against 96.75% for DenseNet121, and two independent groups have proposed transformers with cascaded group attention for tomato [15,16], reporting 96.5% and 96.9% respectively. The evidence that transformers dominate convolutions here is nevertheless weaker than it appears. Where both families are trained under identical conditions, the transformer is not uniformly superior. In the baseline comparison of Ergün [17], a ViT-Base branch reaches only 92.56% where ConvNeXt-Tiny reaches 99.39% on the same data, and Jia et al. [18] report ViT-S below ResNet50 on field imagery (85.33% against 86.08%). The comparison is further confounded by weakly tuned convolutional baselines, and the ResNet-50 reference in [15] collapses to 56.27%. Our own measurements sharpen the point. Across 26 architectures trained under one recipe, no family occupies a contiguous band of the internal ordering. In the reference execution the eight strongest models comprise three transformers, two classical convolutional networks, one modernised convolutional network, one efficiency-oriented network and the proposed hybrid, and the convolutional family spans a wider range of internal macro-F1 (0.9841 down to 0.9677) than the transformer family does (0.9843 down to 0.9814). In our second execution of the same benchmark, the composition of the top eight was different again (Section 4.3).

Hybrid designs are the closest neighbours of the present work, and they differ substantially in where the two inductive biases meet. One family interleaves them within a single stream. ConvTransNet-S [18] stacks blocks combining a local perception unit, lightweight multi-head self-attention and an inverted-residual feed-forward network, reaching 98.85% on PlantVillage and 88.53% on the authors' own 10,441-image field corpus. A second family arranges the two in series. LSeTNet [19] passes squeeze-and-excitation residual convolutional stages into a single transformer encoder block, obtaining 99.72% on a twelve-class medicinal-leaf corpus with 9.38 M parameters, and Murugesan et al. [20] similarly cascade a convolutional stem into a six-layer transformer, reporting 99.29%. A third family runs two backbones in parallel and fuses their outputs. MangoLeafCMD-FAMNet [17] merges ConvNeXt-Tiny and ViT-Base through a squeeze-and-excitation-inspired gate at 46.9 M parameters, LeafFusionNet [21] concatenates a ViT stream with Gabor-filtered convolutional streams, and Pal et al. [22] run MobileNetV2 alongside a residual network at 3.51 M parameters. Hoang and Phan [23] replace fusion with adaptive soft routing between three heterogeneous experts. Our backbone belongs to the first family. CAFormer [24] mixes tokens by separable convolution in its first two stages and by self-attention in its last two, so the two inductive biases are already combined inside one stream and no cross-branch reconciliation is required. What we add sits above the backbone rather than inside it, and, as Section 4.6 reports, none of the additions produces an effect that can be separated from replicate variation at three seeds.

Three limitations recur across the hybrid group. The measured benefit of fusion is often small or measured against weak baselines, and rarely isolated. In [17], fusion alone improves on the stronger single branch by 0.22, 1.46 and 0.55 percentage points on the three corpora and the attention module adds a further 0.17 to 0.28 (differences computed from their reported results), while in [21], the apparent 11.4-point gain is measured against a basic three-layer convolutional network (87.97%) and falls to 5.4 points against their ViT-Base branch (93.9%). With one partial exception, these models are never evaluated outside their training distribution. The exception, [22], evaluates without fine-tuning between PlantVillage and a larger corpus that reuses its label set, and its cross-corpus accuracy is 97.64% when transferring to the simpler corpus but 92.67% when transferring to the harder one. And in [23], the term “cross-dataset” denotes a protocol that retrains separately on each corpus. Improvements of a fraction of a point, reported without any measure of replicate variation and validated only in-domain, are precisely the quantity our replication analysis shows to be uninformative.

Channel attention is near-universal in this literature but placed differently in each case, inside every bottleneck identity block of a custom CNN [25], after each convolutional stage of a VGG16 backbone [26], or on the head of a detector [27]. Reported accuracies are high (99.79% on PlantVillage for CNN-SEEB [25] with only 3.34 M parameters), but attribution of that performance to attention is largely untested. Karthikeyan et al. [25] report their squeeze-and-excitation ablation only as training and validation accuracy curves, without a with/without accuracy value, and Singh et al. [26] never evaluate the backbone without CBAM. In all of these designs, attention recalibrates channels, in [26,27] together with spatial positions. The pooling stage that follows is almost always a global average, which is where our design differs. On a leaf, the great majority of pixels are healthy tissue, so averaging dilutes precisely the evidence the classifier needs. We instead treat the feature map as a bag of patches and pool it with gated attention in the sense of Ilse et al. [28], extended here to multiple heads, a formulation demonstrated on histopathology images but, to our knowledge, not previously applied to leaf disease classification. The cashew model of Meenakshi K and Sankaranarayanan [29], another parallel-branch design, weights its two embeddings with a learned but input-independent fully connected projection and then fits a gradient-boosting classifier on the extracted embeddings, so no gradient from the final classifier reaches the fusion stage.

Almost every model above is trained and tested on a single distribution, and the studies that explicitly measure transfer show how fragile these figures are. Benchmarking across tomato corpora, Sun [4] observed the macro-F1 of a PlantVillage-trained ResNet falling from 0.987 in-domain to 0.012 on Tomato-Village and 0.130 on a merged PlantDoc/Tomato Leaf Disease set, a pattern consistent across ResNet, EfficientNet, Inception and MobileNet, so this is not a defect of any one backbone. Krishna et al. [5] measured the same effect at a milder scale, with four backbones losing 14.0 to 24.0 accuracy points (noise-free setting) moving from a web-sourced corpus to PlantDoc. Yang et al. [8] attribute the laboratory-to-field gap to cluttered backgrounds, uneven illumination and sensor differences, note that fine-grained single-crop identification is harder still, and find PlantVillage to be a worse source domain than insect-pest imagery for cross-domain few-shot learning, which they tentatively attribute to its fixed background. Reviews reach the same conclusion from the corpus side, noting that PlantVillage “faces challenges such as noise, errors, and biases” and that models “struggle” on images from different sensors [7]. Several of these authors point to the same gap. Sun [4] notes that most studies still rely on PlantVillage and that field-collected datasets are scarce and often hard to access, and Yang et al. [8] that existing few-shot learning methods are “trained and evaluated only on a single dataset”. Where cross-dataset protocols are claimed, they are often weaker

than the term implies. PlantCLR [30], published as a generalisable model, fine-tunes on labelled target-domain data before testing and states plainly that its protocol “evaluates cross-domain adaptation rather than zero-shot transfer”. The multi-corpus evaluations in [31,32] likewise train and test within each corpus. Our protocol is the strict version, with identical weights, an identical 21-way head, no target labels and no target statistics.

Grad-CAM is the near-universal choice for justifying model decisions here [10,11,20,33], with occasional use of LIME, layer-wise relevance propagation or attention maps [15,26]. Two gaps are relevant. Explanations are evaluated qualitatively. Across these studies, we found no deletion or insertion curves, no pointing-game scores, no intersection-over-union against lesion annotations and no expert-agreement study. Gopalan et al. [10] list IoU among their evaluation metrics but never compute it, and Alhammad et al. [11] conceded that their heatmaps can be equivocal. Specific to hybrid models, the transformer branch is often not explained at all. The convolution-plus-transformer model of [20] computes its class activation maps on a convolutional layer only. We inherit the first limitation (our own explanation analysis is qualitative, and we say so in Section 5.5 rather than implying otherwise), but the multi-scale fusion stage gives our attribution a 14×14 rather than a 7×7 grid, and the pooling attention supplies a second, gradient-free map over the same features.

Because a design that adds components to a backbone is heavier than the backbone alone, the efficiency literature is directly relevant. Lightweight models have been pushed to genuinely small footprints, namely 2.19 M for SugarcaneShuffleNet [33], 3.34 M for CNN-SEEIB [25], 3.51 M for [22], and 2.13 M for the edge-deployed detector YOLO-PLNet [34], which runs at 11.8 ms and 3.4 W on a Jetson Orin NX at INT8 precision. Al Kafi et al. [35] instead cascade a classical-feature filter ahead of a MobileNet classifier, cutting inference time by up to 77.6% on a low-power laptop and by 46% at a cost of about three points of accuracy in a controlled test. The efficiency–robustness trade-off is not monotone, however. Benchmarking seven architectures on a Raspberry Pi 5, Hoang et al. [31] found them separated by only 1.6 points on laboratory imagery but by 22 points on in-field PlantDoc data, where the 134 M VGG16 scored below the 1.3 M ShuffleNetV2. Commenting on the smallest model’s last place on the laboratory set, they note that “model capacity is still important for capturing fine-grained disease patterns”. Duhan et al. [32] make the mechanism explicit, noting that “lightweight networks generally do not incorporate domain adaptation mechanisms, which diminishes their ability to manage distribution shifts”. In their experiments, a 0.37 M ShuffleNetV2 $\times 0.5$ was 6.6 times faster than MobileNetV2 on a Raspberry Pi 5 but 14.3 points less accurate on PlantDoc field images. Our results reproduce this pattern with 26 models on one protocol, and sharpen it. Internal macro-F1 correlates weakly with parameter count (Spearman’s $\rho = 0.45$, $p = 0.023$), while zero-shot accuracy does not ($\rho = 0.13$, $p = 0.53$) and is better described, over this set, by a concave curve whose shape rests on the two largest networks (Section 4.8). The four smallest models in our set lose between 35 and 54 accuracy points under transfer and the two largest lose 75 to 76.

Table 1 collects the reviewed methods with their evaluation protocols. Its Protocol column is the reason the accuracy column cannot be ranked. Accuracies measured on partitions containing augmented or synthetic copies of training images, or without any held-out test set, are not commensurable with clean held-out results, and the corpora themselves differ in crop, class count and difficulty. The table is also the reason we report replicate-to-replicate variation, since several of the gains claimed in these rows are smaller than the variation we measure in Section 4.3. Of the twenty-five method papers, thirteen address a single crop, six report headline accuracy on partitions containing augmented or synthetic images, three report no held-out test partition, four apply any statistical test, none performs zero-shot evaluation

on an independently collected corpus (one evaluates across two related corpora without fine-tuning), and none assesses explanation quality quantitatively. The present work targets the scope, out-of-distribution and rigour gaps directly, and is deliberately conservative about the fourth. We make no claim of novelty for combining convolution with attention, and no claim of being the most efficient model in the set.

Table 1. Deep learning approaches to leaf disease classification (a contextual summary, not a head-to-head comparison).

Study	Architecture	Crops/Classes	Protocol ^a	Acc. (%)	Par. (M)	XAI	X-DS ^b	Stat. ^c
<i>Hybrid CNN–transformer</i>								
Ergün [17]	ConvNeXt-T + ViT-B, SE fusion	mango, 1/8	CV	99.78	46.9	–	–	✓
Jia et al. [18]	interleaved conv–attention	multi, 14/38	HO	98.85	25.1	–	–	–
<i>same model, field set</i>		multi, 8/39	HO	88.53	25.1	–	–	–
Das et al. [21]	ViT + Gabor-CNN + attention	multi, 14/38	HO	99.33	n/r	G	–	–
Ahmmmed et al. [19]	serial CNN, SE, transformer	medicinal, 3/12	AUG	99.72	9.4	G, L	–	✓ ^d
Murugesan et al. [20]	cascaded ConvNet → ViT	multi, 3/9	HO	99.29	33	G ^e	–	–
<i>Attention/channel recalibration</i>								
Karthikeyan et al. [25]	custom CNN, SE per block	multi, 14/38	HO	99.79	3.3	G	–	–
Singh et al. [26]	VGG16 + CBAM per stage	rice, 1/10	HO	98.87 ⁿ	n/r	G, A, LRP	–	–
<i>Single-branch CNN or ViT</i>								
Hossain et al. [14]	optimised DeiT	mango, 1/8	AUG ^f	99.75	85.8	G	–	–
Ahmmmed et al. [9]	DenseNet201	mango, 1/8	AUG	99.33	20.2	–	–	–
Tuteja et al. [13]	CNN + attention + LSTM	jackfruit, 1/3	AUG	99.00	3.7	G	–	–
Jayaprakash et al. [12]	MobileNetV2, TinyML	cashew, 1/2	AUG ^g	97.70	n/r	–	–	–
Karimanzira [15]	ViT + cascaded group attention	tomato, 1/10	HO	96.50	n/r	DL, A	–	–
Rajendran & Muruganantham [16]	ViT + cascaded group attention	tomato, 1/10	n/r	96.90	n/r	A	–	–
<i>Multi-backbone fusion or routing</i>								
Meenakshi K & Sankaranarayanan [29]	EffNetV2-M + MobileNetV3-S, boosted	cashew, 1/5	HO	93.10 ^h	n/r	–	–	–
Hoang & Phan [23]	3 experts, adaptive soft routing	potato, 1/7	HO ^k	92.28	40.2	G	–	–
Pal et al. [22]	MobileNetV2 + residual CNN	multi, 14/38	HO	99.47 ^l	3.5	G, L	– ¹	–
<i>Explainability-focused</i>								
Arman et al. [33]	ShuffleNet, Bayesian tuning	sugarcane, 1/17	HO ⁱ	98.02	2.2	G	–	–
Gopalan et al. [10]	ResNet152	maize, 1/4	HO	98.34 ^j	n/r	G	–	–
Alhammad et al. [11]	VGG16, frozen backbone	potato, 1/3	AUG	98.00	n/r	G	–	–
This work	CAFormer-S18 + MSF + LAP + crop head, MixStyle	multi, 4/21	HO	98.82	23.9	G, A	✓	✓ ^m
<i>same weights, zero-shot MLD24</i>		<i>mango, 1/6</i>	<i>ZS</i>	<i>69.94</i>	<i>23.9</i>			

Every row reports the value that paper obtained on its own corpus under its own protocol, so the accuracy column must not be read as a ranking. ^a HO denotes a held-out test partition, CV cross-validation only with no held-out test, AUG an evaluation partition that contains augmented or synthetic images, ZS zero-shot evaluation on an independently collected corpus, and n/r not reported. ^b Genuine zero-shot evaluation, with no fine-tuning and no target-domain statistic. ^c Any significance testing of the reported differences. ^d Reported, but the paired tests compare cross-validation folds against single test values, and the authors' own Friedman test is not significant ($p = 0.540$). ^e Computed on a convolutional layer only. The transformer pathway is not explained. ^f The source corpus (MangoLeafBD) already contains zoomed and rotated copies (about 1800 distinct leaves among 4000 images), and the split is random. The test partition holds 400 images, so 99.75% corresponds to one error. ^g GAN-synthesised images enter the evaluation data. On the original images alone, the same model gives 96%. ^h Single-split value. The paper's own five-fold mean is 91.90 ± 0.32 . ⁱ Cleanest protocol in the set, with MD5 de-duplication ($9908 \rightarrow 7037$), a stratified split and augmentation restricted to the training partition. ^j The text accompanying the test confusion matrix reports 93.69% for the same experiment. ^k Test partition of 311 images (7 in the smallest class), single seed, no significance testing. ^l Measured on the validation split of PlantVillage (no separate test partition). Mean of three training strategies rather than a single run. The proposed model is timed at 128×128 while all baselines are evaluated at 224×224 . Cross-dataset evaluation: trained on one corpus and evaluated on the other without fine-tuning, but the second corpus reuses the PlantVillage label set. ^m The Friedman and Nemenyi procedures, McNemar's test, and the Holm-corrected Wilcoxon procedures are reported, and none is significant (see Section 4.2). The tick records that the tests were run and their outcome disclosed, not that a difference was established. ⁿ Highest of the five corpora evaluated (Apple, Embrapa, Maize, PlantVillage, Rice), which give 94.20–98.87%. G = Grad-CAM, L = LIME, DL = DLIME, A = attention maps, LRP = layer-wise relevance propagation. Detection studies [27,34,36] are excluded because they report mean average precision (mAP), which is not commensurable with classification accuracy. Italics mark the model-family headings and the additional rows for the same model (field set) or the same weights (zero-shot on MLD24); bold marks this work. A check mark (✓) indicates that the item is present, and – that it is absent.

3. Materials and Methods

Figure 1 summarises the whole experimental procedure. All 47 configurations (25 generic backbones, the proposed model, 16 single-component ablation variants and five backbones retrained with the proposed model's complete recipe) share the same partitions, the same implementation and the same training schedule, with the settings specified per configuration in Sections 3.5 and 3.6, so that any difference between them is attributable to the configuration and not to the data they received. The 26 generic-recipe configurations were, in addition, executed twice from scratch, three weeks apart on the same machine. We call these executions replication A and replication B. Replication A, the first execution, is the reference execution reported in Section 4 unless stated otherwise. Replication B is reported in Section 4.3, where the pair is used to measure how much of any difference is reproducible, and it also supplies the recipe-matched, repeated and symmetry-group experiments.

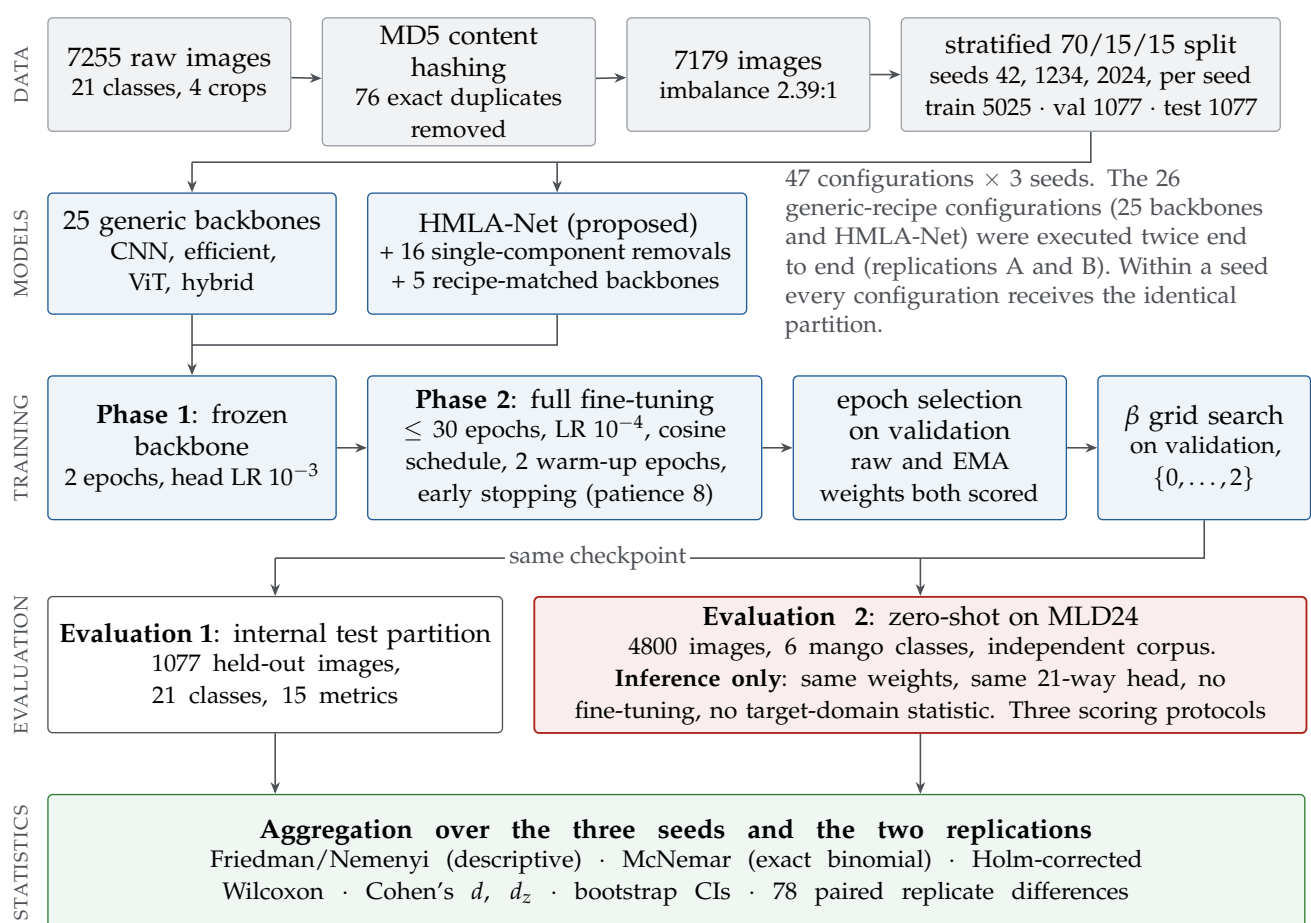


Figure 1. Experimental pipeline. Within a seed, every one of the 47 configurations receives the identical stratified partition, which is what makes the paired comparisons of Section 3.8 valid. The two evaluations consume the same checkpoint, and Evaluation 2 adds no training step, no target labels and no target-domain statistic of any kind. The 26 generic-recipe configurations (25 backbones and HMLA-Net) were executed twice end to end. Replication A, the first execution, is the reference reported in the main tables, and the A–B pair is analysed in Section 4.3.

3.1. Corpora

The internal corpus is the Multi-Crop Leaf Disease Dataset for Deep Learning-Based Classification and Explainable AI [37], which contains 7255 images of 480×360 pixels organised as 21 classes across four crops, namely mango (6 classes), tomato (7), jackfruit (6)

and cashew (2). We removed 76 exact duplicates by MD5 content hashing before any partitioning, leaving 7179 images. This matters because near-identical images straddling a train/test boundary inflate held-out accuracy without any modelling error being detectable. The resulting class sizes range from 209 (mango sooty mould) to 500, an imbalance ratio of 2.39:1. Table 2 gives the composition and Figure 2 the per-class distribution. Figure 3 shows one image per class and makes an acquisition property visible that will matter later. The cashew, jackfruit and mango images were captured against a near-white background and the tomato images against a bluish-grey one, so crop identity is partly encoded in the background.

Table 2. Composition of the four-crop corpus.

Crop	Classes	Images	Range	Disease Classes
Cashew	2	480	235–245	fungal infection, healthy
Jackfruit	6	1814	220–368	algal, anthracnose, healthy, leaf spot, phyllosticta, sooty mold
Mango	6	2688	209–500	anthracnose, die back, gall midge, healthy, powdery mildew, sooty mould
Tomato	7	2197	222–394	bacterial spot, blight, curl virus, healthy, insect damage, leaf mold, spider mites
Total	21	7179	209–500	

Counts are after content-hash (MD5) de-duplication, which removed 76 exact duplicates from the 7255 raw files. The imbalance ratio is 2.39:1. Each class of the distributed corpus is a folder named by crop and condition inside the folder of that crop, and the crop is the label that the auxiliary crop head of Section 3.3 predicts. Class names are quoted as the corpus defines them (it writes “jackfruit sooty mold” and “tomato leaf mold”), whereas the text uses “mould”.

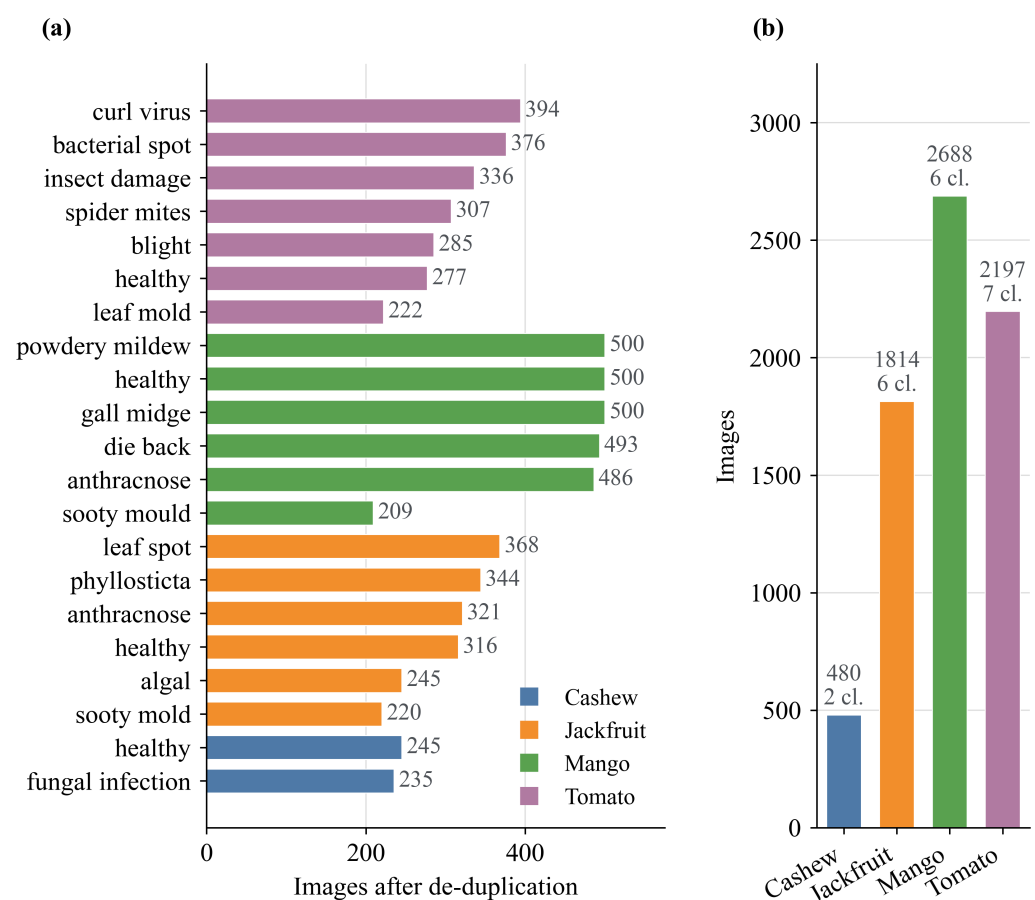


Figure 2. Composition of the internal corpus after de-duplication. (a) Images per class, coloured by crop. (b) Images and class count per crop. The corpus is moderately imbalanced (2.39:1) and unevenly distributed across crops, which is why macro-averaged metrics are used throughout and why a class-balanced sampler is part of the common training recipe.

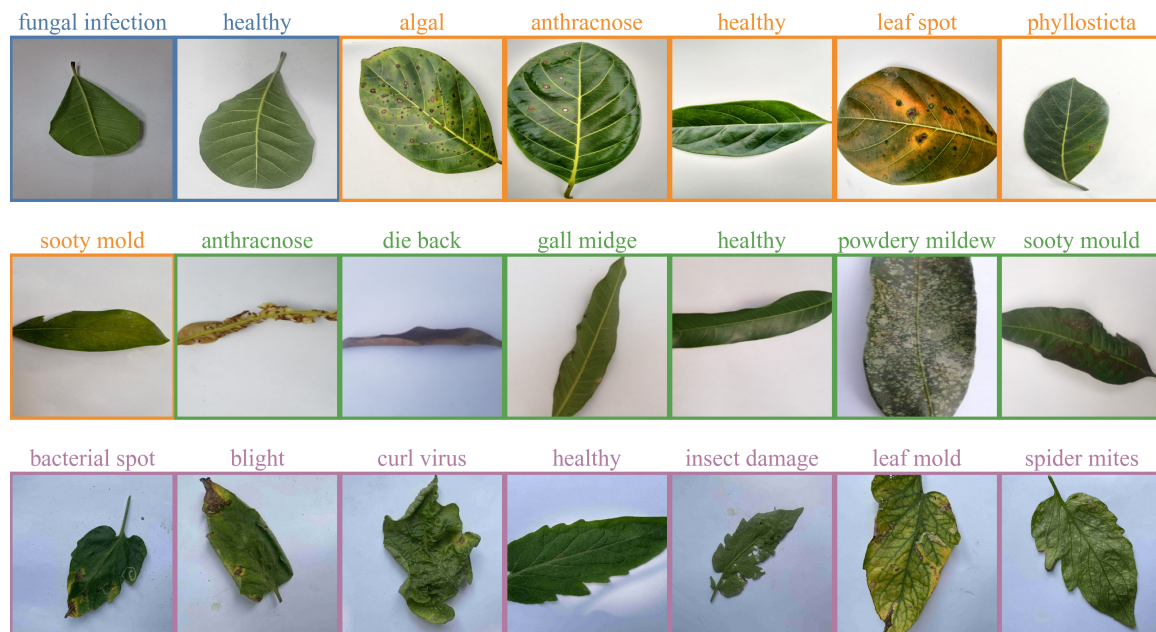


Figure 3. One representative image from each of the 21 classes, drawn directly from the distributed corpus. Frame and caption colour encode the crop (blue for cashew, orange for jackfruit, green for mango and purple for tomato). Lesion scale varies by more than an order of magnitude between classes. Jackfruit phyllosticta is a scatter of millimetric spots while mango die back affects the whole leaf, which is the observation the multi-scale fusion stage of Section 3.3 is built on. The tomato images also carry a visibly different background tint from the other three crops, a nuisance correlate that MixStyle is intended to suppress.

For external validation, we use MLD24 [38], an independently assembled mango corpus of 6400 images in eight classes of 800 each, captured with an iPhone SE and distributed at 240×240 pixels. Six of its classes map onto labels in our space (anthracnose, die back, gall midge, healthy, powdery mildew, sooty mould). The remaining two, bacterial canker and cutting weevil, have no counterpart in our corpus and are excluded, leaving 4800 evaluation images. Both corpora originate in Bangladesh and both are CC BY 4.0, so this is not a cross-geography test. The shift is one of acquisition and curation, with a different research group, a different capture session, a different camera, and a different delivered resolution (240×240 against 480×360). Because the source corpus covers four crops but MLD24 covers mango only, every cross-corpus statement in this paper is a statement about six mango classes. Cross-corpus behaviour for tomato, jackfruit and cashew is not measured here, and we repeat this in the abstract, the discussion and the conclusion so that the title is not read more broadly than the evidence.

Duplicate audit between the corpora.

Because a zero-shot protocol is only as strong as the independence of its two corpora, we checked for overlap in three ways. Exact matching by MD5 content hash finds no image of MLD24 in the internal corpus. Perceptual hashing (64-bit pHash, Hamming distance ≤ 6) flags 117 of the 4800 MLD24 images (2.4%) as having a near neighbour in the internal corpus. Inspecting the listed pairs shows what this detector responds to on plain-background leaf photographs. Of the 50 pairs archived with the report, 4 join images of the same class, 29 join an MLD24 image to a different mango class and 17 to a different crop altogether (a jackfruit or tomato leaf), and 22 internal images act as hubs matched to many MLD24 images, so these are structural coincidences of a centred leaf on a light background rather than copies. The mean nearest-neighbour pHash distance is 12.3 out of 64 bits, and only 15 pairs lie at distance ≤ 4 . Finally, a comparison of ImageNet-pretrained

embeddings finds one MLD24 image with cosine similarity ≥ 0.95 to any internal image (maximum 0.952, 99th percentile 0.902). The pair joins a healthy MLD24 leaf to a healthy internal leaf and was not verified as a copy by inspection, so we count it as a candidate. We conclude that no exact duplicates were detected and that the screening provides no evidence of substantial overlap. Thresholds of this kind do not have perfect sensitivity, so near-copies below them cannot be excluded, and the low transfer scores of Section 4.4 corroborate but do not by themselves establish the independence of the two corpora.

3.2. Partitioning

Within each seed $s \in \{42, 1234, 2024\}$, the de-duplicated corpus is split by stratified sampling into 70% training (5025 images), 15% validation (1077) and 15% test (1077). The split is computed once per seed and reused by every configuration, so the three seeds constitute three paired blocks for the Friedman and Wilcoxon procedures [39], although, unlike the independent data sets that framework assumes, they are correlated (Section 3.8). The partition files and their SHA-256 fingerprints are provided in the public repository (Data Availability Statement). The test partition is touched only for final reporting, since model selection, early stopping and the calibration of the head-combination coefficient all use the validation partition.

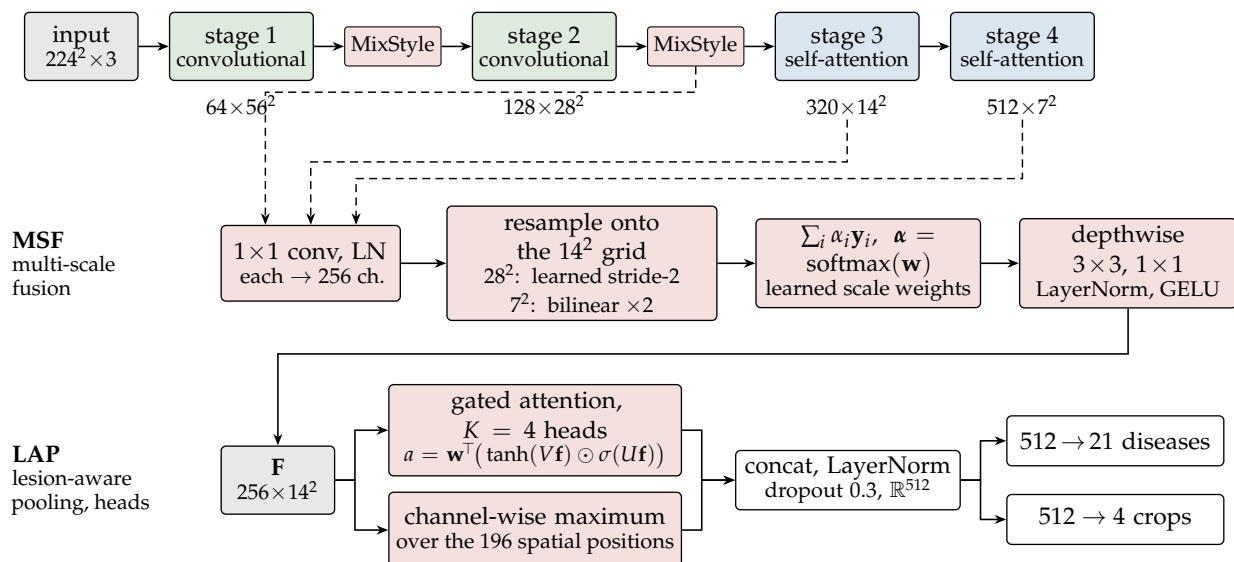
The same perceptual hash applied within the internal corpus shows that the split is not near-duplicate-aware. Our implementation of the 64-bit pHash (a median-thresholded 8×8 DCT of a 32×32 grey image, slightly more inclusive than the audit tool of the previous paragraph, with 1781 against 1423 pairs at the same threshold) finds 302 to 383 test–training pairs at Hamming distance ≤ 6 per seed, marking 186 to 240 of the 1077 test images (17 to 22%) as having a training neighbour. Of the pairs, roughly two-thirds join images of the same class and a few (8 in seed 42) are hash-identical burst shots. To measure what this does to the internal scores, every run of the reference execution was re-scored on the clean test subset that excludes all flagged images (870, 837 and 891 images in the three seeds). The result is the opposite of inflation, since 25 of the 26 architectures score higher on the clean subset, by 0.17 pp of macro-F1 on average (-0.11 to $+0.43$ pp) and 0.12 pp of accuracy, because the flagged images are harder than the rest (mean accuracy 0.980 on the flagged images against 0.985 on the clean subset). The between-model spread widens slightly from 1.68 to 1.85 pp, the ordering is essentially unchanged (Spearman's $\rho = 0.91$ between full and clean macro-F1, largest rank move seven places), and HMLA-Net moves from 0.9845 to 0.9869 and remains first. The same re-scoring of the second execution gives the same picture (all 31 of its configurations rise, by 0.20 pp on average). Excluding the flagged test images therefore does not reduce the internal scores of this benchmark, so the saturation of Section 4.1 does not depend on them. This sensitivity analysis covers the test partition only. Near-duplicates between the training and validation partitions may have influenced checkpoint selection, and a duplicate-group-aware partition would be the cleaner design (Section 5.5). The per-configuration values of both executions are provided in the repository.

3.3. Proposed Architecture

HMLA-Net is shown in Figure 4, whose caption also states where each component sits in the network. The backbone is CAFormer-S18 [24] pretrained on ImageNet-22k and fine-tuned on ImageNet-1k [40], a MetaFormer instance whose first two stages mix tokens by separable convolution and whose last two use self-attention. We take the backbone as given rather than as a contribution. The bare backbone is included as one of the 25 baselines, and it is also retrained with the complete recipe of the proposed model (Section 3.5). The effect of everything we add can therefore be read off under identical training and inference conditions.

CAFormer-S18 backbone (ImageNet-22k $\rightarrow$ 1k, where stages 1–2 mix tokens by convolution, stages 3–4 by self-attention)

MixStyle acts on the output of stages 1 and 2, and only during training: batch feature statistics (μ, σ) are mixed across samples, $\lambda \sim \text{Beta}(0.1, 0.1)$, with probability 0.5. At inference, it is the identity, so it costs nothing at deployment.



At inference the heads are combined as $\ell_c = \ell_c^{\text{dis}} + \beta \log p(\text{crop}_c)$, with β selected on validation from a nine-value grid between 0 and 2. The grid contains 0, so the combination can never harm. On all three seeds $\beta = 0$ was selected.

Figure 4. HMLA-Net and the location of each component. The backbone is CAFormer-S18 (top band, where stages 1–2 mix tokens by separable convolution and stages 3–4 by self-attention). MixStyle wraps the outputs of stages 1 and 2 (feature maps of 64×56^2 and 128×28^2) and is active only in training, where it mixes per-sample feature statistics within the batch. At inference, it is the identity. MSF (middle band) takes the outputs of stages 2, 3 and 4, projects each to 256 channels, resamples them onto the stride-16 grid (14×14) and combines them with learned scale weights. Its output $\mathbf{F}$ is the Grad-CAM target. LAP (bottom band) pools the 196 positions of $\mathbf{F}$ with gated multi-head attention in the multiple-instance sense, concatenated with a channel-wise maximum, into a 512-dimensional descriptor. The pooled descriptor feeds the 21-way disease head and the 4-way auxiliary crop head, whose combination coefficient β is calibrated on validation. Tensor shapes are given for a 224×224 input. The total parameter count is 23.9 M, which is 0.4 M fewer than the bare backbone with its 1000-way ImageNet classifier replaced by a 21-way one. Box fills encode the type of block: green, convolutional backbone stages; blue, self-attention backbone stages; light red, components added by HMLA-Net; grey, input and feature tensors; white, descriptor normalisation and classification heads.

Multi-scale fusion (MSF).

Figure 3 shows that lesion scale varies by more than an order of magnitude across our classes, and a single stride-32 map cannot carry that range. Let $\mathbf{x}_1 \in \mathbb{R}^{128 \times 28^2}$, $\mathbf{x}_2 \in \mathbb{R}^{320 \times 14^2}$ and $\mathbf{x}_3 \in \mathbb{R}^{512 \times 7^2}$ be the outputs of backbone stages 2, 3 and 4. Each is projected to a common width $d = 256$ by a 1×1 convolution followed by channel-axis layer normalisation, and then resampled onto the stride-16 grid. The finer map is reduced by a learned depthwise stride-2 convolution rather than average pooling, and the coarser one is upsampled bilinearly. We write $\mathbf{y}_i$ for the map after both steps, so that all three carry the shape $256 \times 14 \times 14$. The choice of a learned reduction is deliberate, because average pooling a 28×28 map to 14×14 averages a millimetric lesion pixel with its three healthy neighbours, which would defeat the purpose of consulting the fine scale at all. The

three maps are then combined with learned, softmax-normalised scalar weights in the manner of [41] rather than summed uniformly as in a feature pyramid [42], because the dominant scale differs by class:

$$\mathbf{F} = \phi\left(\sum_{i=1}^3 \alpha_i \mathbf{y}_i\right), \quad \alpha = \text{softmax}(\mathbf{w}), \quad \mathbf{w} \in \mathbb{R}^3, \quad (1)$$

In Equation (1), ϕ is a depthwise 3×3 convolution followed by a pointwise 1×1 convolution, layer normalisation and GELU. The output $\mathbf{F} \in \mathbb{R}^{256 \times 14 \times 14}$ is also the Grad-CAM target, so attribution maps are computed at 14×14 instead of the 7×7 a stride-32 backbone would offer.

Lesion-aware attentive pooling (LAP).

Lesions typically occupy a small fraction of the leaf area, so global average pooling divides the lesion signal by the area of healthy tissue. We instead treat the $N = 196$ spatial positions of $\mathbf{F}$ as a bag of instances and pool them with the gated attention formulation of Ilse et al. [28]. For position n with feature $\mathbf{f}_n \in \mathbb{R}^{256}$ and head $k \in \{1, \dots, K\}$, $K = 4$, as follows:

$$a_{k,n} = \frac{1}{\tau} \mathbf{w}_k^\top \left(\tanh(V\mathbf{f}_n) \odot \sigma(U\mathbf{f}_n) \right), \quad \alpha_{k,n} = \frac{\exp(a_{k,n})}{\sum_{m=1}^N \exp(a_{k,m})}, \quad (2)$$

with $V, U \in \mathbb{R}^{128 \times 256}$, a learned temperature τ constrained to $[0.2, 5]$, and $\odot$ the Hadamard product. The sigmoid gate is what distinguishes this from a plain tanh scorer, which is approximately linear for inputs in $[-1, 1]$ and may therefore limit the expressiveness of the learned attention [28]; the gate adds a learnable non-linearity. The gate modulates how much each patch contributes to the attention score separately from the feature content it carries. Multiple heads allow different spatial weighting patterns to be learned, which matters because a leaf may carry several distinct lesion appearances at once, for example necrosis alongside a chlorotic halo. The heads are concatenated and projected back to d , so the descriptor width does not depend on K , and are concatenated with a channel-wise maximum that preserves the single strongest response:

$$\mathbf{z} = \left[W_p \left(\oplus_{k=1}^K \sum_n \alpha_{k,n} \mathbf{f}_n \right) \parallel \max_n \mathbf{f}_n \right] \in \mathbb{R}^{512}. \quad (3)$$

In Equation (3), $\oplus$ denotes concatenation over the K heads, $W_p \in \mathbb{R}^{256 \times 1024}$ is the projection that restores the width, and $\parallel$ separates the two branches. The attention maps $\alpha_{k,n}$ are retained at inference and averaged over heads, giving a gradient-free saliency map that is independent of Grad-CAM.

Auxiliary crop head and calibrated combination.

Every class name carries a crop prefix, so the 21 diseases form four disjoint groups. After layer normalisation and dropout ($p = 0.3$), $\mathbf{z}$ feeds a disease head $\mathbf{f}_{\text{dis}} : \mathbb{R}^{512} \rightarrow \mathbb{R}^{21}$ and a crop head $\mathbf{f}_{\text{plant}} : \mathbb{R}^{512} \rightarrow \mathbb{R}^4$. The crop head is trained with a cross-entropy term of weight $\lambda = 0.3$ added to the disease loss. At inference the two are combined as

$$\ell_c = \ell_c^{\text{dis}} + \beta \log p(\text{crop}(c)), \quad c = 1, \dots, 21, \quad (4)$$

which leaves the output dimension unchanged. Because the loss is computed on ℓ^{dis} alone while inference uses Equation (4), a fixed β introduces a train/inference mismatch, in which a confident but wrong crop prediction suppresses every disease of that crop simultaneously. We therefore treat β as a post hoc parameter and select it on the validation partition over the grid $\{0, 0.1, 0.2, 0.3, 0.5, 0.75, 1, 1.5, 2\}$ by macro-F1, using a single forward pass whose

raw logits are rescored for every grid point. The grid contains zero, so the selected value cannot underperform $\beta = 0$ on the validation selection metric. Ties are resolved in favour of the first grid value, which is zero. We report in Section 4.6 that zero is in fact what the calibration selects on all three seeds, and we discuss what that means.

MixStyle.

As Figure 3 shows, the tomato images differ from the other three crops in background tint, and background is exactly the kind of cue that does not survive a change of acquisition setup. MixStyle [43] attacks this directly by mixing first- and second-order feature statistics between samples in a batch. Wrapping the outputs of the two convolutional stages (style information is concentrated in shallow layers), we replace $\mathbf{x}$ with

$$\hat{\mathbf{x}} = (\lambda\sigma(\mathbf{x}) + (1 - \lambda)\sigma(\mathbf{x}_\pi)) \frac{\mathbf{x} - \mu(\mathbf{x})}{\sigma(\mathbf{x})} + (\lambda\mu(\mathbf{x}) + (1 - \lambda)\mu(\mathbf{x}_\pi)), \quad (5)$$

In Equation (5), μ and σ are per-sample, per-channel spatial statistics, π is a random permutation of the batch, and $\lambda \sim \text{Beta}(0.1, 0.1)$. It is applied with probability 0.5 during training only, and is an identity map at inference, so it costs nothing at deployment.

3.4. Training Protocol

Algorithm 1 states the procedure that every configuration follows. Training has two phases. The backbone is frozen for two epochs while the newly initialised head is brought up to a usable state at learning rate 10^{-3} , then the whole network is fine-tuned for up to 30 epochs at 10^{-4} with AdamW [44], weight decay 10^{-4} excluded from normalisation weights, biases and learned scalars, a two-epoch linear warm-up followed by a cosine schedule [45], and early stopping after eight epochs without improvement. Layer-wise learning-rate decay [46] with $\gamma = 0.75$ scales the rate down towards the shallow stages. Augmentation is random-resized crop (scale 0.6–1.0), horizontal and vertical flips, $\pm 30^\circ$ rotation, colour jitter and random erasing [47]. A class-balanced sampler addresses the 2.39:1 imbalance, and label smoothing [48] of 0.1 is applied throughout. MixUp [49] and CutMix [50] are used with probability 0.8 and switched off for the last eight epochs so the network can settle on the unmixed distribution it will be tested on.

Symmetry of the acquisition geometry and orbit averaging.

Test-time augmentation exploits a symmetry of the acquisition geometry. A detached leaf photographed against a plain background has no canonical orientation, so the disease label is invariant under horizontal reflection, under vertical reflection, and under their composition. Together with the identity, these four maps form the Klein four-group $V_4 \cong \mathbb{Z}_2 \times \mathbb{Z}_2$, the subgroup of the dihedral group D_4 generated by the two axis reflections. Two distinct things are done with this group, and they should not be conflated. During training, the augmentation draws uniformly from V_4 , because the horizontal and vertical flips are applied independently with probability 1/2 each. This only encourages the network to be approximately equivariant and gives no guarantee, and we measure how far the trained network actually is from invariance in Section 4.9. At inference, we average the posterior over the whole group orbit,

$$p(c | x) = \frac{1}{|V_4|} \sum_{g \in V_4} p_\theta(c | g \cdot x), \quad (6)$$

so that $p(c | h \cdot x) = p(c | x)$ for every $h \in V_4$ as an algebraic identity. The prediction becomes exactly invariant under the group rather than approximately so, at the cost of four forward passes. Four-view flip averaging is a well-known technique and we claim

no novelty for it. What this paper contributes is the formal statement of the constraint as a group orbit, the verification that the orbit average is invariant to floating-point precision, and the measurement of what the constraint costs and buys under distribution shift, including a comparison with the trivial group and with the full dihedral group D_4 (Section 4.9).

Two details of Algorithm 1 are worth isolating because they are easy to get wrong. First, exponential moving averaging of the weights is a standard stabiliser [51,52], but its decay must be set from the number of optimisation steps, not copied from an ImageNet recipe. With roughly 158 steps per epoch, a decay of 0.9998 gives an effective window of 5000 steps, the entire run, so a third of the shadow weights would still be the randomly initialised head. We derive the decay from a three-epoch horizon and apply a warm-up correction $(1+t)/(10+t)$. Second, the raw and the averaged weights are both evaluated on validation every epoch and the checkpoint is taken from whichever scores higher, at a cost of about 3% of run time. Blindly trusting the shadow copy is the failure mode the first detail describes, and evaluating both lets the selection keep whichever candidate validates better.

Algorithm 1 Training and evaluation of one configuration (one seed). All configurations, including every baseline and every ablation variant, execute this procedure through the same implementation. The steps that belong to optional elements (MixUp/CutMix, EMA weights, the crop head and its β calibration, orbit averaging) run only when the configuration enables them, so only the model constructor and the flags it reads differ.

Input: de-duplicated corpus $\mathcal{D}$; seed s ; configuration $\mathcal{C}$; external corpus $\mathcal{E}$; label map $m : \mathcal{E} \rightarrow \{1, \dots, 21\}$

Output: internal test metrics, zero-shot metrics on $\mathcal{E}$

```

1:  $(\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{va}}, \mathcal{D}_{\text{te}}) \leftarrow \text{STRATIFIEDSPLIT}(\mathcal{D}, 0.70/0.15/0.15, s)$   $\triangleright$  computed once per seed, shared by
   all configurations
2: seed every random number generator with  $s$ ; build  $M \leftarrow \mathcal{C}(\text{pretrained})$ 
3:  $M_{\text{ema}} \leftarrow \text{COPY}(M)$ ;  $d \leftarrow \min(0.9999, 1 - 1/(3|\mathcal{D}_{\text{tr}}|/B))$   $\triangleright$  decay from step count, not a
   constant
4:  $b^* \leftarrow -\infty$ ;  $u \leftarrow 0$ 
5: for  $e = 0$  to  $E_{\text{freeze}} + E_{\text{ft}} - 1$  do
6:   if  $e < E_{\text{freeze}}$  then freeze the backbone, optimise the head at  $10^{-3}$  else unfreeze all, optimise
     at  $10^{-4}$  with layer-wise decay  $\gamma$ 
7:   for each mini-batch  $(\mathbf{X}, \mathbf{y})$  of  $\mathcal{D}_{\text{tr}}$  do
8:     if MixUp active then  $(\mathbf{X}, \mathbf{y}) \leftarrow \text{MIXUPCUTMIX}(\mathbf{X}, \mathbf{y})$   $\triangleright$  disabled for the final  $E_{\text{cool}}$  epochs
9:      $(\ell^{\text{dis}}, \ell^{\text{plant}}) \leftarrow M(\mathbf{X})$   $\triangleright$  MixStyle active, Equation (4) not applied in training
10:     $\mathcal{L} \leftarrow \text{CE}_{\text{ls}}(\ell^{\text{dis}}, \mathbf{y}) + \lambda \text{CE}(\ell^{\text{plant}}, \mathbf{y}\Pi)$   $\triangleright \Pi$ : class-to-crop matrix, so mixed targets stay
       consistent
11:    update  $M$  (AdamW, AMP);  $M_{\text{ema}} \leftarrow d M_{\text{ema}} + (1 - d) M$  with warm-up correction
12:  end for
13:   $q \leftarrow \max(\text{F1}_{\text{macro}}(M, \mathcal{D}_{\text{va}}), \text{F1}_{\text{macro}}(M_{\text{ema}}, \mathcal{D}_{\text{va}}))$   $\triangleright$  both candidates scored, every epoch
14:  if  $q > b^*$  then  $b^* \leftarrow q$ ; checkpoint the better candidate;  $u \leftarrow 0$  else  $u \leftarrow u + 1$ 
15:  if  $e \geq E_{\text{freeze}}$  and  $u \geq 8$  then break
16: end for
17:  $\beta \leftarrow \arg \max_{\beta' \in \mathcal{B}} \text{F1}_{\text{macro}}(\ell^{\text{dis}} + \beta' \log \text{softmax}(\ell^{\text{plant}}), \mathcal{D}_{\text{va}})$ ,  $0 \in \mathcal{B}$   $\triangleright$  one forward pass, grid
   rescored on the logits
18: return  $\text{EVAL}(M^*, \mathcal{D}_{\text{te}}, \text{TIA})$ ,  $\text{EVAL}(M^*, m(\mathcal{E}), \text{TIA})$   $\triangleright M^*$  unchanged between the two, no
   adaptation
```

3.5. Baselines and the Recipe-Matched Comparison

Twenty-five ImageNet-pretrained backbones from timm [53] span the design space, covering classical convolutional networks (ResNet-50/101 [54], DenseNet-121/201 [55], VGG-16/19 [56], Inception-v3 [48], Xception [57], ResNeXt-50 [58], SE-ResNeXt-50 [59]), efficiency-oriented networks (MobileNetV2 [60], MobileNetV3-L [61], EfficientNet-B0/B3 [62], MnasNet-B1 [63], RegNetY-1.6GF [64], GhostNet [65]), modernised

convolutional networks (ConvNeXt-T/S [66]), vision transformers (ViT-S/16 [67,68], ViT-B/16 [67], Swin-T, Swin-S [69], DeiT III-S [68]), and the bare CAFormer-S18 backbone [24].

All 25 receive the common recipe (two-phase schedule, class-balanced sampler, random erasing, label smoothing, identical augmentation, identical optimiser and schedule), but not the components specific to the proposed model (MSF, LAP, the crop head, MixStyle) nor the four recipe elements that are ablated as part of the proposal (MixUp/CutMix, EMA, orbit-averaged TTA, layer-wise decay). A comparison between HMLA-Net and these 25 rows therefore compares an architecture plus an extended recipe against backbones under the generic one, and cannot by itself attribute anything to the architecture. To make the attribution possible, five of the strongest backbones (CAFormer-S18, EfficientNet-B3, Swin-S, ViT-S/16 and SE-ResNeXt-50) were retrained with the complete HMLA-Net training and inference recipe, MixUp/CutMix, EMA weights, layer-wise decay and V_4 orbit averaging included, with the architectural components switched off. A machine-readable recipe matrix is provided in the repository. The results of these five recipe-matched configurations are reported separately. The comparison between HMLA-Net and recipe-matched CAFormer-S18 is the one on which any claim about the added components must rest, and Section 4.5 reports it together with the replicate variation against which it must be read.

3.6. Ablation Design

Sixteen variants each remove one element from the full model and are retrained from scratch with the same three seeds and the same partitions. Together with the full model, they form 17 configurations. Five removals are architectural (MSF, LAP, LAP gating together with multi-head attention, the auxiliary crop head, MixStyle) and eleven belong to the training recipe, one of which removes ImageNet pretraining entirely. Each variant is measured on both the internal test partition and MLD24, and the whole ablation was executed in both replications, so every component carries two independent estimates of its effect on each partition. All 16 removals completed all three seeds on both partitions in replication B. In replication A, the MixUp cool-down variant had completed two of three seeds on the internal partition, and that replication's value for it is reported with that caveat. The pretraining variant is reported in the tables but excluded from the figures because its effect is an order of magnitude larger than everything else and would compress the rest of the scale.

3.7. Metrics and Zero-Shot Scoring Protocols

Fifteen metrics are computed per run. Let a partition hold N images with reference labels y_i , predictions $\hat{y}_i$ and predicted probability vectors $\mathbf{p}_i \in \mathbb{R}^{21}$, and let M be the confusion matrix, $M_{cd} = |\{i : y_i = c, \hat{y}_i = d\}|$. For class c , we write $\text{TP}_c = M_{cc}$, $\text{FP}_c = \sum_{d \neq c} M_{dc}$, $\text{FN}_c = \sum_{d \neq c} M_{cd}$ and $\text{TN}_c = N - \text{TP}_c - \text{FP}_c - \text{FN}_c$, and define

$$\begin{aligned} \text{Acc} &= \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{y}_i = y_i], & \text{Pre}_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \\ \text{Rec}_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, & \text{Spe}_c &= \frac{\text{TN}_c}{\text{TN}_c + \text{FP}_c}, \end{aligned} \quad (7)$$

with $\text{F1}_c = 2 \text{Pre}_c \text{Rec}_c / (\text{Pre}_c + \text{Rec}_c)$ completing the per-class scores of Equation (7).

Let $\mathcal{C}$ be the set of classes that occur in the reference labels of the partition being scored, so $|\mathcal{C}| = 21$ internally and $|\mathcal{C}| = 6$ on MLD24. Every macro average is taken over $\mathcal{C}$ alone:

$$X_{\text{macro}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} X_c, \quad X \in \{\text{Pre}, \text{Rec}, \text{Spe}, \text{F1}\}, \quad \text{BAcc} = \text{Rec}_{\text{macro}},$$

$$G = \exp\left(\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \ln \max(\text{Rec}_c, \varepsilon)\right), \quad (8)$$

with $\varepsilon = 10^{-12}$ in Equation (8). Because $|\mathcal{C}|$ differs between the two partitions, the macro scores of the internal and the external tables are of different cardinality (21 against 6 classes) and are not directly comparable to each other. The comparison the paper makes is always between models on the same partition, never between partitions of the same model except as a degradation. On MLD24 the head still emits 21 logits, so a model may assign an image to any of the 15 labels that MLD24 does not contain. We score this in three protocols so that the choice of denominator cannot hide anything:

- **P1**, the primary protocol, takes the argmax over all 21 logits and macro-averages over the six reference classes $\mathcal{C}$. A prediction into an absent label counts as a false negative for its true class and is otherwise ignored.
- **P2** uses the same predictions, macro-averaged over the full 21-label output space. Because an absent class has no reference images, its F1 is zero whether or not it is ever predicted, so $\text{F1}_{\text{P2}} = \frac{6}{21} \text{F1}_{\text{P1}}$ exactly. P2 therefore carries no information beyond P1, and we report it only to make that identity explicit.
- **P3**, a separate experiment, restricts the logits to the six shared classes before the argmax, so that no prediction can leak. This is the score a practitioner would obtain by telling the model the crop, and it is not zero-shot in the strict sense.

Alongside these we report the leakage rate, the fraction of MLD24 images assigned to one of the 15 absent labels under P1. It is the quantity that distinguishes P1 from P3, and it is itself informative, because a model that routes mango leaves into tomato or jackfruit classes is making a meaningful error that a six-class denominator would not show. The floor ε in G plays a similar role for the recall side. Without it, a single class at zero recall would send the geometric mean to exactly 0 for every affected model, erasing the differences between them. With it, one zero recall drives G below 0.01 whatever the remaining recalls ($\varepsilon^{1/6} = 0.01$ times the geometric mean of the other five), which is what makes the column a direct witness of the single-class collapse analysed in Section 4.7.

Writing $t_c = \sum_d M_{cd}$ for the reference count of class c and $q_c = \sum_d M_{dc}$ for how often it was predicted, chance-corrected agreement is measured by Cohen's κ and by the multiclass Matthews correlation coefficient of Equation (9),

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad p_o = \text{Acc}, \quad p_e = \frac{1}{N^2} \sum_c t_c q_c,$$

$$\text{MCC} = \frac{N \sum_c M_{cc} - \sum_c t_c q_c}{\sqrt{(N^2 - \sum_c q_c^2)(N^2 - \sum_c t_c^2)}}. \quad (9)$$

Ranking quality is summarised by the one-versus-rest area under the ROC curve, $\text{AUC}_c = \Pr(p_{ic} > p_{jc} \mid y_i = c, y_j \neq c)$, and by average precision, both macro-averaged over $\mathcal{C}$ and both computed on the raw probability column $p_{\cdot c}$ rather than on probabilities renormalised over the covered classes, since a per-row renormalisation constant that varies by model would change a rank-based statistic.

Calibration is measured by the expected calibration error of Equation (10) over $B = 15$ equal-width bins of the top-class confidence $\hat{p}_i = \max_c p_{ic}$:

$$\text{ECE} = \sum_{b=1}^B \frac{|\mathcal{B}_b|}{N} \left| \frac{1}{|\mathcal{B}_b|} \sum_{i \in \mathcal{B}_b} \mathbf{1}[\hat{y}_i = y_i] - \frac{1}{|\mathcal{B}_b|} \sum_{i \in \mathcal{B}_b} \hat{p}_i \right|, \quad (10)$$

where $\mathcal{B}_b$ collects the images whose confidence falls in the b -th bin [70]. Macro-F1 is the primary metric throughout and is also the model-selection criterion of Algorithm 1, so selection and reporting are consistent.

3.8. Statistical Analysis and Replicate Variation

We follow Demšar [39] for the seed-level comparisons, with two caveats stated in advance. First, that framework was developed for comparing classifiers across multiple independent datasets. Three random partitions of one corpus are not independent datasets, and the seed blocks are correlated through the images they share, so the Friedman and Nemenyi statistics are reported here as descriptive summaries of the rank structure and not as inferential tests. Second, the design is underpowered by construction. With three paired blocks, the smallest two-sided p value the Wilcoxon signed-rank test can attain is $2^{1-N} = 0.25$. Reaching $\alpha = 0.05$ would need at least six blocks, and surviving Holm correction over the 25 comparisons of Table 3 would need ten (eleven for the 30 comparisons among the 31 configurations of the second execution). These counts are the points at which rejection becomes arithmetically possible in the most favourable case, not a power calculation. The Nemenyi critical difference is $q_\alpha \sqrt{k(k+1)/(6N)}$, which, for $k = 26$, $N = 3$ and $\alpha = 0.05$, is 22.97 average ranks on an axis that is only 25 ranks wide. Consequently, the statement that “no pair of models can be separated” is a statement about the design, not about the models. It is not independent evidence that the models are equivalent, and we keep it separate from the empirical observation that the spread between models is comparable to the spread between replicates of one model. The computational budget allowed three seeds per configuration for 47 configurations. We report the tests because they are the field’s convention, and we report their attainable power beside them.

The proposed model is compared against each baseline with a paired Wilcoxon signed-rank test corrected by the Holm–Bonferroni [71], and against the strongest baseline with McNemar’s test [72,73] on the individual test predictions. Because the number of discordant pairs between two strong models is routinely below 25, we use the exact binomial form, which cannot exceed the nominal level, rather than the χ^2 approximation, whose level is not guaranteed in this regime; like the continuity-corrected statistic, the exact form is conservative, which works against our own model. Effect sizes are reported both as the paired d_z and as the pooled Cohen’s d . On correlated scores measured over shared splits, d_z is much the larger of the two, and reporting only that quantity under the label “Cohen’s d ” inflates the apparent effect. Accuracy confidence intervals come from 2000 bootstrap resamples of the test predictions.

The quantity that does carry information at this budget is the replicate-to-replicate difference. Every one of the 26 generic-recipe configurations was trained twice on every seed, in two executions of the whole benchmark separated by three weeks on the same machine, giving 78 paired differences per partition and metric. HMLA-Net was in addition trained a third time on every seed (as the ablation reference) and CAFormer-S18 with the matched recipe a second time, giving nine and six runs respectively of those two configurations. We report the full distribution of these differences (mean absolute value, standard deviation, maximum and the rank correlation between the two orderings), and we use them as the yardstick against which every between-model and between-variant difference in this paper is read. We deliberately avoid the term “noise floor” for this

quantity, because a finite set of repeat runs estimates the typical size of replicate variation, not its lower bound.

Table 3. Held-out internal test performance of the 26 architectures on the four-crop, 21-class corpus. The whole benchmark spans 1.68 pp of macro-F1 and the top eight models span 0.11 pp; Section 4.2 shows that no pair of models in this table can be separated statistically.

#	Model	Family	Accuracy	Macro-F1	Bal. acc.	MCC	AUC	ECE	Par. (M)	GFLOPs
1	HMLA-Net (ours)	hybrid	0.9882 ± 0.0014	0.9845 ± 0.0012	0.9847	0.9876	0.9998	0.079	23.9	3.97
2	Swin-S	ViT	0.9882 ± 0.0030	0.9843 ± 0.0045	0.9844	0.9876	0.9992	0.081	48.9	8.77
3	VGG-16	CNN	0.9873 ± 0.0030	0.9841 ± 0.0039	0.9841	0.9867	0.9995	0.097	134.3	15.47
4	SE-ResNeXt-50	CNN	0.9879 ± 0.0009	0.9840 ± 0.0008	0.9841	0.9873	0.9994	0.077	25.6	4.26
5	RegNetY-1.6GF	eff.	0.9873 ± 0.0035	0.9837 ± 0.0043	0.9836	0.9866	0.9995	0.127	10.3	1.63
6	ConvNeXt-S	CNN*	0.9876 ± 0.0035	0.9835 ± 0.0044	0.9838	0.9870	0.9991	0.092	49.5	8.70
7	Swin-T	ViT	0.9870 ± 0.0056	0.9834 ± 0.0061	0.9837	0.9863	0.9994	0.085	27.5	4.51
8	ViT-B/16	ViT	0.9870 ± 0.0028	0.9834 ± 0.0027	0.9835	0.9863	0.9995	0.093	85.8	16.87
9	DenseNet-201	CNN	0.9861 ± 0.0019	0.9827 ± 0.0020	0.9828	0.9853	0.9993	0.100	18.1	4.34
10	ViT-S/16	ViT	0.9861 ± 0.0046	0.9824 ± 0.0054	0.9825	0.9853	0.9995	0.095	21.7	4.25
11	DenseNet-121	CNN	0.9864 ± 0.0048	0.9823 ± 0.0055	0.9822	0.9857	0.9998	0.100	7.0	2.86
12	DeiT III-S	ViT	0.9855 ± 0.0030	0.9814 ± 0.0034	0.9814	0.9847	0.9995	0.093	21.7	4.25
13	CAFormer-S18 [†]	hybrid	0.9855 ± 0.0033	0.9812 ± 0.0044	0.9811	0.9847	0.9993	0.096	24.3	3.89
14	Inception-v3	CNN	0.9848 ± 0.0035	0.9811 ± 0.0047	0.9813	0.9841	0.9994	0.097	21.8	2.85
15	ConvNeXt-T	CNN*	0.9848 ± 0.0062	0.9803 ± 0.0082	0.9802	0.9841	0.9997	0.091	27.8	4.47
16	MobileNetV3-L	eff.	0.9839 ± 0.0030	0.9797 ± 0.0031	0.9793	0.9831	0.9997	0.127	4.2	0.22
17	VGG-19	CNN	0.9839 ± 0.0028	0.9797 ± 0.0043	0.9801	0.9831	0.9995	0.091	139.7	19.63
18	EfficientNet-B3	eff.	0.9839 ± 0.0046	0.9796 ± 0.0062	0.9794	0.9830	0.9984	0.128	10.7	0.99
19	Xception	CNN	0.9839 ± 0.0033	0.9794 ± 0.0039	0.9797	0.9831	0.9992	0.257	20.9	4.57
20	ResNeXt-50	CNN	0.9830 ± 0.0044	0.9791 ± 0.0045	0.9792	0.9821	0.9995	0.107	23.0	4.26
21	EfficientNet-B0	eff.	0.9830 ± 0.0023	0.9786 ± 0.0038	0.9788	0.9821	0.9995	0.106	4.0	0.40
22	MnasNet-B1	eff.	0.9814 ± 0.0025	0.9772 ± 0.0023	0.9775	0.9805	0.9997	0.111	3.1	0.32
23	GhostNet	eff.	0.9808 ± 0.0051	0.9758 ± 0.0065	0.9754	0.9798	0.9996	0.126	3.9	0.15
24	MobileNetV2	eff.	0.9759 ± 0.0019	0.9706 ± 0.0026	0.9709	0.9746	0.9994	0.090	2.3	0.31
25	ResNet-50	CNN	0.9762 ± 0.0023	0.9704 ± 0.0031	0.9706	0.9749	0.9995	0.129	23.6	4.11
26	ResNet-101	CNN	0.9737 ± 0.0030	0.9677 ± 0.0023	0.9681	0.9723	0.9994	0.101	42.5	7.83

Mean ± standard deviation over three seeds (42, 1234, 2024); within a seed every model receives the identical 70/15/15 partition. Models are ordered by macro-F1. eff. = efficiency-oriented CNN; CNN* = modernised CNN (ConvNeXt); hybrid = interleaved convolution and attention. [†] The bare backbone of the proposed model, included so that the contribution of the added components can be separated from the contribution of the backbone. ECE = expected calibration error (15 bins, lower is better); AUC is macro one-versus-rest. Complexity is measured at 224 × 224 with a batch of one. The row of the proposed model is set in bold.

3.9. Explainability

Grad-CAM [74] is computed at the MSF output, a 14 × 14 grid. The LAP attention weights of Equation (2), averaged over the four heads, provide a second saliency map that requires no gradients and is derived from a different mechanism. Class structure in the 512-dimensional pooled space is visualised with t-SNE [75].

3.10. Implementation and Reproducibility

Experiments ran on an NVIDIA RTX A4000 under Windows 10 with Python 3.11.5, PyTorch 2.5.1 (CUDA 12.1, cuDNN 9.1), timm 1.0.28, scikit-learn 1.9.0 and SciPy 1.17.1, using mixed precision. Every run writes its complete configuration, per-epoch history, predictions and metrics. Replication A (128 runs) required approximately 23.5 GPU hours of training and replication B (150 runs, including the recipe-matched and repeat runs) approximately 57.7, excluding evaluation and analysis. Because cuDNN autotuning is left enabled, runs are not bit-wise reproducible even at a fixed seed. Rather than

assume this variability away, we measure it, and the two-execution design of Section 3.8 is how. The configuration of every run, the recipe matrix, the partition files with their SHA-256 fingerprints, the archived metrics and per-image predictions of both replications, and the scripts that produce every table and figure from them are published together with the training and evaluation code in a public repository, <https://github.com/songulkarakus/HMLA-Net> (version 1.1.2; accessed on 6 September 2026), with a manifest and reproduction instructions.

4. Results

Unless stated otherwise, the tables and figures of this section report the reference execution of the benchmark, the first of the two (replication A). The second, independent execution (replication B) is reported in Section 4.3, where the two are compared configuration by configuration. That execution also supplies the recipe-matched baselines of Section 4.5, the repeated trainings, and the symmetry-group evaluation of Section 4.9. The reference was fixed by chronology, not by outcome, and every difference between the two executions is reported.

4.1. The Internal Benchmark Is Saturated

Table 3 and Figure 5 report the held-out internal test results. HMLA-Net attains the highest macro-F1 of the 26 architectures in this execution, 0.9845 ± 0.0012 , and the highest accuracy, 0.9882 ± 0.0014 , with balanced accuracy 0.9847, MCC 0.9876 and macro AUC 0.9998, at 23.9 M parameters and 3.97 GFLOPs.

These numbers should not be read as a ranking, and the leaderboard position of HMLA-Net is a fact about this table rather than a property of the model. The 26 architectures span 1.68 percentage points of macro-F1, from 0.9677 (ResNet-101) to 0.9845. The gap from the first to the second model is 0.017 pp and from the first to the eighth 0.11 pp. Against this, Section 4.3 shows that retraining the same configuration on the same seed and partition moves its macro-F1 by 0.34 pp on average per run and by up to 1.06 pp, that the three-seed mean the table reports moves by 0.19 pp on average, and that in the second execution the same HMLA-Net configuration is eighth. Figure 5 draws those two widths as bands below the top of the table, and 13 configurations lie inside the mean band and 23 inside the maximum band. The difference between most of these models is smaller than the difference between one configuration and itself, and the re-scoring of Section 3.2 shows that this is not an artefact of near-duplicate images in the test partition.

4.2. What the Seed-Level Tests Can and Cannot Say

The Friedman statistic over all 26 models is nominally significant ($\chi^2 = 40.98$, $p = 0.023$), but the accompanying Nemenyi critical difference is 22.97 average ranks while the entire observed range of average ranks is 19.33, from 6.67 to 26.0. Figure 6 shows the critical-difference bar exceeding the axis it is drawn on. Restricting the analysis to the eight strongest models removes even the nominal effect ($\chi^2 = 0.33$, $p = 0.9999$). As Section 3.8 anticipated, this is a property of the design and not of the models. With three paired blocks, the smallest attainable Wilcoxon p is 0.25, and the Holm-corrected tests against the 25 baselines reject nothing, with the raw p values of the closest comparisons sitting exactly at that floor. McNemar's exact test against the strongest baseline, Swin-S, gives $p = 0.727$, 0.227 and 0.375 on the three seeds with 8, 11 and 5 discordant pairs out of 1077 test images, and the bootstrap 95% confidence intervals for HMLA-Net's accuracy, [0.9805, 0.9944], [0.9796, 0.9935] and [0.9833, 0.9954], contain every baseline in the top eight. We draw one conclusion from these tests and refuse a second. The conclusion is that seed-level testing at this budget cannot order the leading models, which is why we do not

describe any in-domain difference in this paper as significant. What we do not conclude is that the models are therefore equivalent. That would be reading the absence of power as evidence, and the empirical case for saturation rests instead on the replicate analysis that follows.

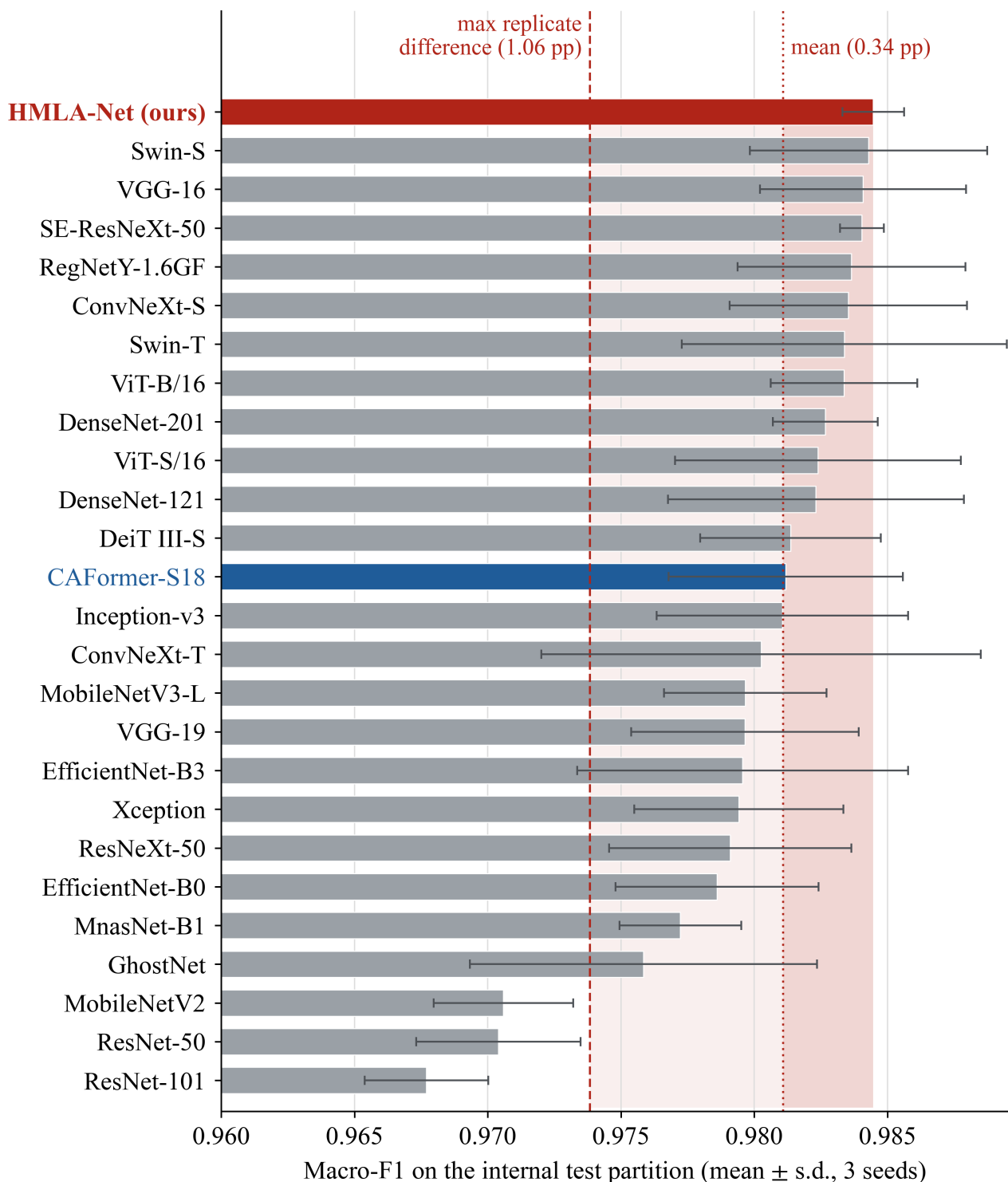


Figure 5. Internal test macro-F1 for the 26 architectures of the reference execution (mean $\pm$ s.d. over three seeds). The two shaded bands below the best value are the mean (0.34 pp) and the maximum (1.06 pp) replicate-to-replicate difference of a single configuration measured over 78 paired runs in Section 4.3, and 13 configurations lie inside the first band and 23 inside the second. HMLA-Net is drawn in red and its bare backbone in blue.

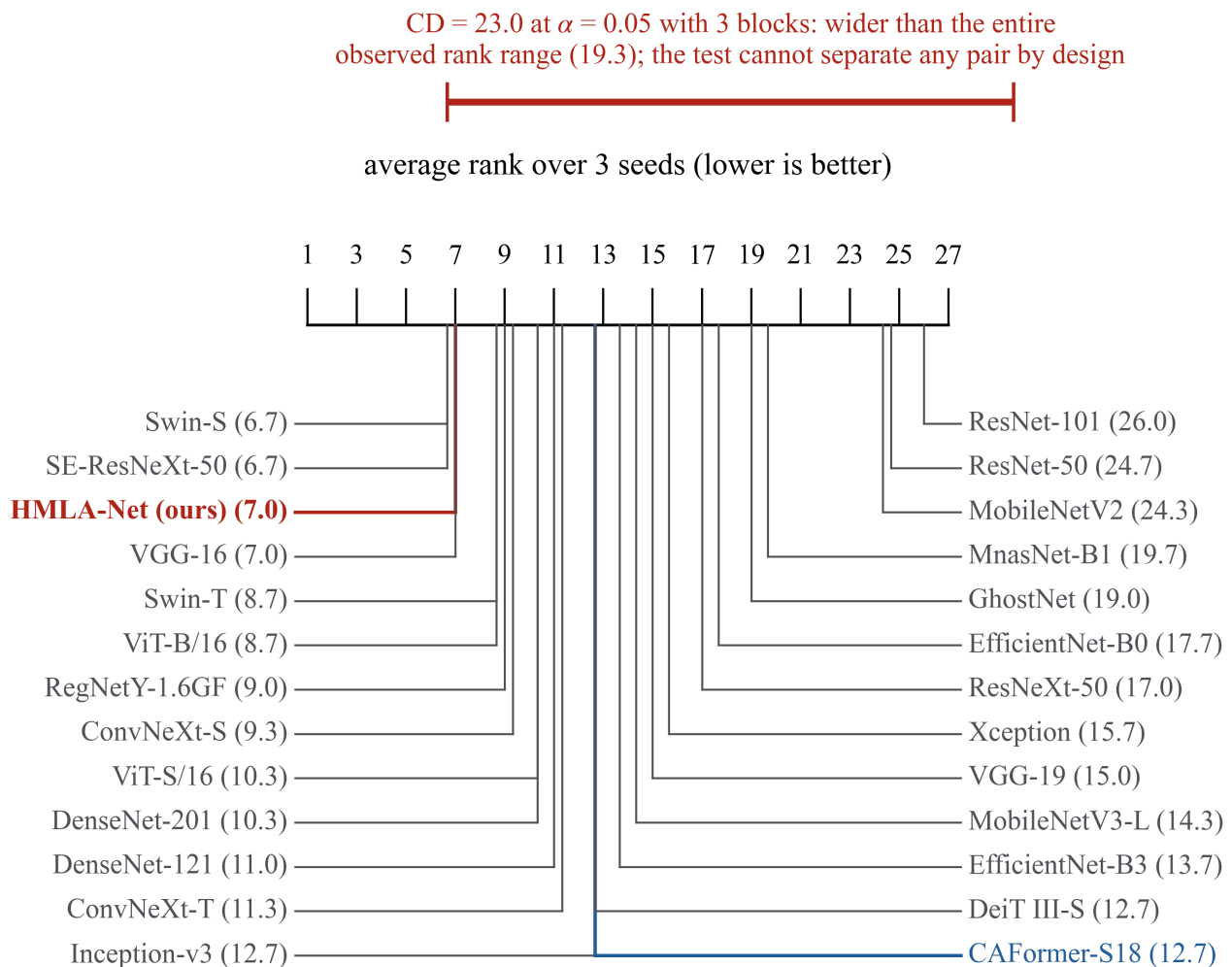


Figure 6. Critical-difference diagram for the internal test partition of the reference execution (Friedman’s test with Nemenyi’s post hoc procedure, $\alpha = 0.05$, three blocks, 26 models). The critical difference of 23.0 average ranks exceeds the observed range of 19.3, so every model is connected to every other and the diagram’s usual function, separating groups, is unavailable by design. We reproduce it in this form deliberately, because the same test applied to the same number of seeds in the comparable literature carries the same limitation.

4.3. Two Executions of the Benchmark: What Replicates and What Does Not

Table 4 and Figure 7 compare the reference execution with the second, independent execution of the same 26 configurations. On the internal partition the same configuration on the same seed and split differs between executions by 0.34 pp of macro-F1 on average (median 0.27, standard deviation 0.42, 90th percentile 0.69, maximum 1.06 pp), the three-seed means that the tables report differ by 0.19 pp on average and by 0.72 pp at most, and the two orderings agree only partly, with Spearman’s $\rho = 0.69$, and eleven of the 26 configurations move five or more places. The largest moves are VGG-19 from 17th to 1st, ConvNeXt-S from 6th to 17th, ResNeXt-50 from 20th to 9th and HMLA-Net from 1st to 8th. The between-model spread (1.68 pp in A, 1.72 pp in B) is about five times the mean replicate difference, which is enough to keep the bottom of the table apart from the top but not enough to order the upper two-thirds. This is the observation the paper’s title rests on, and it is an empirical one that does not depend on any significance test.

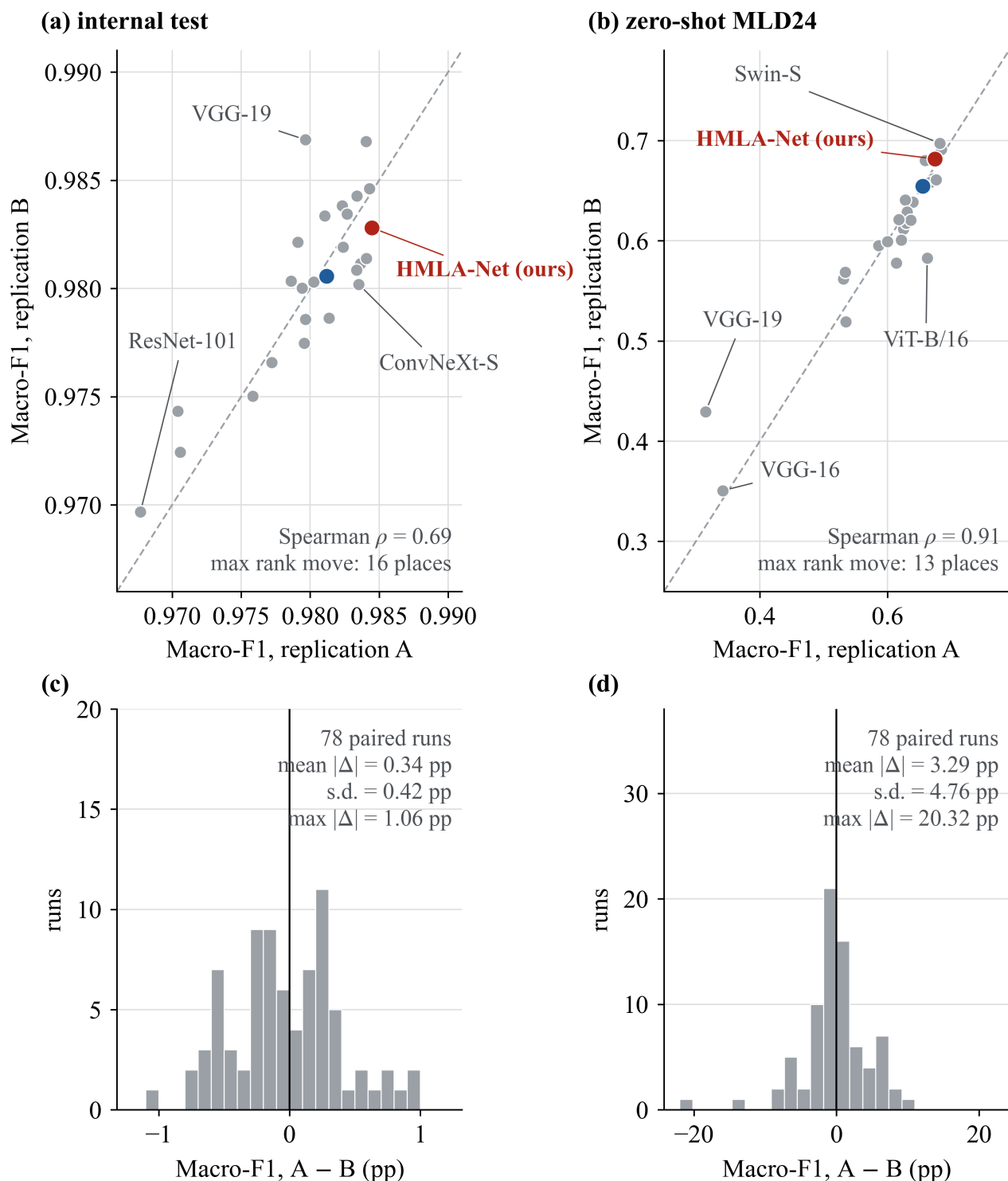


Figure 7. Two end-to-end executions of the same 26-configuration benchmark. **(a,b)** Three-seed mean macro-F1 of each configuration in the reference execution (A) against the second execution (B), on the internal test partition and under zero-shot transfer. The dashed line is identity. HMLA-Net is drawn in red and its bare CAFormer-S18 backbone in blue. **(c,d)** Distribution of the 78 paired differences (same configuration, seed and partition). The internal ordering replicates only partly ($\rho = 0.69$, with VGG-19 moving 16 places) although its absolute variation is small. The external ordering replicates well ($\rho = 0.91$) although its absolute variation is ten times larger, because the between-model spread is larger still.

Table 4. Replicate-to-replicate variation measured by re-executing the whole benchmark.

Partition, Metric	<i>n</i>	Mean $ \Delta $	s.d.(Δ)	Max $ \Delta $	Mean $ \bar{\Delta} $	Max $ \bar{\Delta} $	Spread A	Spread B	ρ_{AB}	Max Rank Move
internal test, macro-F1	78	0.34	0.42	1.06	0.19	0.72	1.68	1.72	0.69	16
internal test, accuracy	78	0.29	0.36	0.93	0.17	0.59	1.45	1.49	0.70	14
zero-shot MLD24, macro-F1	78	3.29	4.76	20.32	1.91	11.42	36.90	34.67	0.91	13
zero-shot MLD24, accuracy	78	4.10	5.57	19.17	2.29	10.54	47.29	44.03	0.92	11

Two complete, independent executions of the 26 generic-recipe configurations (25 backbones and the proposed model, three seeds each), carried out three weeks apart with the same code, recorded hyper-parameters, partitions and hardware. Non-deterministic cuDNN kernels and data-loader ordering are the expected sources of the difference. Δ is A minus B for the same configuration, seed and partition, in percentage points, and s.d.(Δ) is the standard deviation of the signed Δ . $\bar{\Delta}$ is A minus B for the three-seed means of a configuration, the quantity on which the tables of this paper compare models. The spread columns give the range of the 26 three-seed means, ρ_{AB} is the Spearman correlation between the two orderings, and the max rank move column is the largest change in rank of any configuration. HMLA-Net itself was trained three times in replication B (benchmark entry, ablation reference and a dedicated repeat), and its nine runs give internal macro-F1 0.9837 ± 0.0041 and external macro-F1 0.6601 ± 0.0293 , with the three-seed means of the replicates at 0.6818, 0.6345, and 0.6639 externally.

Under zero-shot transfer, the picture inverts. The absolute replicate difference is ten times larger (mean 3.3 pp of macro-F1, median 2.0, standard deviation 4.8, maximum 20.3 pp for VGG-19), 21 of the 78 paired runs differ by more than five points, and the three-seed means differ by 1.9 pp on average and by 11.4 at most. Yet the two orderings agree closely ($\rho = 0.91$ for macro-F1 and 0.92 for accuracy, mean rank move 2.0 places), because the between-model spread of 35 to 37 pp of macro-F1 is more than ten times the replicate difference. The one configuration that moves substantially, ViT-B/16 from 7th to 20th, does so with a mean replicate difference of 7.9 pp, the third largest in the set after the two VGG networks. HMLA-Net is first in zero-shot accuracy in both executions and, with a mean replicate difference of 0.94 pp of accuracy, the most stable of the 26 configurations on that metric. Both partitions therefore carry the same lesson in different forms. The external partition orders the families of models reproducibly and the internal partition does not, and neither partition, at three seeds, orders the models within its leading group.

HMLA-Net itself was trained three times in the second execution, as the benchmark entry, as the reference of the ablation and as a dedicated repeat, in addition to the two trainings of the reference execution (benchmark entry and ablation reference). Internally, the three replicates of the second execution agree to within 0.40 pp on every seed (three-seed means 0.9828, 0.9838 and 0.9845, nine-run mean 0.9837 ± 0.0041). Externally, they do not. The three-seed means of macro-F1 are 0.6818, 0.6345 and 0.6639 (nine-run mean 0.6601 ± 0.0293), the same seed differs between replicates by up to 7.3 pp, and the accuracy means are 0.7040, 0.6531 and 0.6832, whereas the reference execution's two trainings gave 0.6739 and 0.6751. The standard deviation across the three seeds of one execution, ± 0.0043 for the second execution's benchmark entry, understates the nine-run standard deviation of the configuration (± 0.0293) by a factor of about seven (the other two replicates give ± 0.0237 and ± 0.0337), which is a warning about seed-level error bars under distribution shift, because they measure the dispersion of one execution, not the dispersion of the configuration. Recipe-matched CAFormer-S18, trained twice, behaves the same way (0.6531 and 0.6262 externally, paired differences of 2.1, 0.9 and 5.1 pp). We use the pooled nine-run and six-run means wherever a difference involving these two configurations is interpreted below.

4.4. Zero-Shot Transfer Separates Families That the Internal Partition Cannot

Applying the same 26 checkpoints to MLD24, with the 21-way head unchanged and no adaptation of any kind, produces Table 5 and Figure 8. Because MLD24 contains six classes, every macro-averaged quantity in Table 5 is computed over those six classes and is therefore of a different cardinality from the 21-class macro averages of Table 3 (Section 3.7). The two tables are compared model against model, not partition against partition. The 26 architectures now span 47.3 percentage points of accuracy, from 0.2265 (VGG-19) to 0.6994, and 36.9 points of macro-F1, from 0.3152 to 0.6842. The 1.68-point in-domain spread has become a 47-point out-of-domain spread over the same weights.

The ordering also changes. Spearman's ρ between internal macro-F1 and external accuracy is 0.39 ($p = 0.052$). Between the internal and external macro-F1 rankings, it is 0.51 ($p = 0.008$), and Kendall's τ between rank vectors is 0.36 ($p = 0.010$). The relationship is real but weak, and it is not monotone in the region that matters. VGG-16, third of 26 in-domain, falls to twenty-fifth with 0.242 accuracy, and VGG-19 follows it. Conversely, EfficientNet-B3, eighteenth in-domain, is second out-of-domain, and, as Section 4.3 showed, the internal ordering is in any case not stable enough to predict anything.

HMLA-Net ranks first of 26 on external accuracy (0.6994 ± 0.0116 , with the next being EfficientNet-B3 at 0.6926) and fourth on external macro-F1 (0.6739 ± 0.0225), behind SE-ResNeXt-50 (0.6842), Swin-S (0.6819) and ViT-S/16 (0.6759). The five leading configurations lie within 1.33 pp of macro-F1 of one another, which is less than half the mean replicate difference on this partition. We therefore describe HMLA-Net as a member of the leading group, and we do not rank the members of that group by macro-F1. Section 4.7 explains why the accuracy and macro-F1 ranks differ, and the explanation is informative rather than incidental. HMLA-Net also shows the smallest degradation of any model in the set, losing 28.9 points of accuracy between the partitions against a median of 36.1 and a maximum of 75.7 (VGG-19). Its first place on accuracy and its smallest degradation are the two facts about it that the second execution reproduces. Because the internal accuracy of Table 3 is measured over 21 classes and the MLD24 accuracy over six, this difference mixes the change of label space with the change of corpus. Restricted to the same six mango classes, the internal accuracy of HMLA-Net is 1.000 (18 of the 26 architectures reach 1.000 and none falls below 0.9975), so the like-for-like loss is 30.1 points against a median of 37.9, the two orderings coincide ($\rho = 0.998$), and HMLA-Net keeps the smallest loss. The per-model values are provided in the repository. Against its own bare backbone under the generic recipe, it gains 7.1 points of accuracy and 1.9 of macro-F1, while the backbone alone ranks twelfth externally. That comparison confounds architecture with recipe, and Section 4.5 makes the matched comparison.

Table 5. Zero-shot transfer to MLD24 (4800 images, six mango classes, an independent corpus collected in Bangladesh). The same 26 checkpoints span 47 pp of accuracy and 37 pp of macro-F1, and the internal ranking predicts the external one only weakly (Spearman's $\rho = 0.39$, $p = 0.052$).

#	# _{int}	Model	Accuracy	Macro-F1	F1 _{restr.}	Leak (%)	AUC	ECE	G-mean	Δ acc. (pp)
1	1	HMLA-Net (ours)	0.6994 $\pm$ 0.0116	0.6739 $\pm$ 0.0225	0.6758	7.4	0.9204	0.083	0.251	28.9
2	18	EfficientNet-B3	0.6926 $\pm$ 0.0145	0.6709 $\pm$ 0.0056	0.6767	7.7	0.9077	0.119	0.256	29.1
3	10	ViT-S/16	0.6863 $\pm$ 0.0261	0.6759 $\pm$ 0.0191	0.6784	12.5	0.9083	0.051	0.121	30.0
4	4	SE-ResNeXt-50	0.6735 $\pm$ 0.0228	0.6842 $\pm$ 0.0131	0.7054	16.3	0.9251	0.033	0.246	31.4
5	5	RegNetY-1.6GF	0.6642 $\pm$ 0.0261	0.6670 $\pm$ 0.0123	0.6872	14.5	0.9162	0.076	0.195	32.3
6	2	Swin-S	0.6635 $\pm$ 0.0068	0.6819 $\pm$ 0.0068	0.6969	19.4	0.9160	0.064	0.288	32.5
7	8	ViT-B/16	0.6598 $\pm$ 0.0436	0.6619 $\pm$ 0.0437	0.6745	11.0	0.9022	0.130	0.482	32.7
8	6	ConvNeXt-S	0.6432 $\pm$ 0.0482	0.6395 $\pm$ 0.0336	0.6543	14.2	0.8984	0.141	0.008	34.4
9	26	ResNet-101	0.6430 $\pm$ 0.0183	0.6361 $\pm$ 0.0291	0.6427	13.4	0.9077	0.055	0.414	33.1

Table 5. Cont.

#	# _{int}	Model	Accuracy	Macro-F1	F1 _{restr.}	Leak (%)	AUC	ECE	G-mean	Δacc. (pp)
10	9	DenseNet-201	0.6322 ± 0.0625	0.6587 ± 0.0492	0.6904	20.8	0.9180	0.099	0.008	35.4
11	22	MnasNet-B1	0.6288 ± 0.0363	0.6304 ± 0.0221	0.6500	11.1	0.9080	0.154	0.389	35.3
12	13	CAFormer-S18 [†]	0.6284 ± 0.0679	0.6548 ± 0.0429	0.6989	21.8	0.9316	0.111	0.281	35.7
13	20	ResNeXt-50	0.6251 ± 0.0176	0.6175 ± 0.0126	0.6439	14.7	0.9157	0.054	0.087	35.8
14	15	ConvNeXt-T	0.6172 ± 0.0144	0.6247 ± 0.0123	0.6450	17.7	0.9129	0.131	0.168	36.8
15	25	ResNet-50	0.6124 ± 0.0131	0.6212 ± 0.0104	0.6440	17.4	0.8985	0.108	0.195	36.4
16	7	Swin-T	0.6042 ± 0.0115	0.6520 ± 0.0117	0.6818	28.0	0.9155	0.062	0.279	38.3
17	23	GhostNet	0.5965 ± 0.0562	0.6277 ± 0.0511	0.6735	23.1	0.9189	0.135	0.337	38.4
18	19	Xception	0.5945 ± 0.0381	0.5858 ± 0.0445	0.5960	15.6	0.9038	0.159	0.007	38.9
19	11	DenseNet-121	0.5656 ± 0.0320	0.6291 ± 0.0206	0.7200	33.8	0.9205	0.118	0.320	42.1
20	12	DeiT III-S	0.5620 ± 0.0205	0.6137 ± 0.0105	0.6759	31.6	0.9055	0.105	0.007	42.3
21	16	MobileNetV3-L	0.5532 ± 0.0419	0.6001 ± 0.0445	0.6411	28.7	0.8811	0.153	0.213	43.1
22	14	Inception-v3	0.4908 ± 0.0471	0.5339 ± 0.0384	0.5913	30.2	0.8794	0.049	0.173	49.4
23	24	MobileNetV2	0.4881 ± 0.0706	0.5348 ± 0.0437	0.5953	29.8	0.8657	0.084	0.074	48.8
24	21	EfficientNet-B0	0.4482 ± 0.0214	0.5311 ± 0.0182	0.6411	38.9	0.8637	0.080	0.333	53.5
25	3	VGG-16	0.2417 ± 0.0895	0.3422 ± 0.1012	0.5923	69.4	0.8013	0.233	0.003	74.6
26	17	VGG-19	0.2265 ± 0.0184	0.3152 ± 0.0364	0.6022	72.0	0.7786	0.327	0.073	75.7

Zero-shot: The checkpoints of Table 3 are applied to MLD24 without any training, fine-tuning or target-domain statistic. Because MLD24 covers 6 of the 21 labels, every macro average here is taken over those six classes and is therefore not of the same cardinality as the 21-class macro averages of Table 3 (Section 3.7). Models are ordered by accuracy, and #_{int} is the rank of the same model in Table 3. F1_{restr.} is macro-F1 when the logits are restricted to the six shared classes (protocol P3, a separate experiment). Leak is the share of MLD24 images assigned to one of the 15 labels absent from MLD24. Δacc. is the accuracy lost between the partitions. Because the six classes are equally sized, accuracy, balanced accuracy and macro-recall coincide. G-mean is the geometric mean of the six per-class recalls, so one class at zero recall drives it below 0.01 whatever the remaining recalls. The replicate-to-replicate difference of the same configuration on this partition is 3.3 pp of macro-F1 on average (Section 4.3), so the leading group is not internally ordered. The proposed model is set in bold; [†] denotes the bare backbone of the proposed model.

Where the leaked predictions go.

Under the primary protocol, a model may assign an MLD24 leaf to any of the 15 labels the corpus does not contain, and the Leak column of Table 5 reports how often each does. The rate ranges from 7.4% for HMLA-Net (the lowest of the 26 architectures) through a median of 18.5% to 72.0% for VGG-19. Among the other members of the leading group, EfficientNet-B3 leaks 7.7%, ViT-S/16 12.5%, SE-ResNeXt-50 16.3% and Swin-S 19.4%, and the bare CAFormer-S18 backbone 21.8%. Leakage is the mechanism behind the gap between accuracy and macro-F1 in the leading group. A leaked prediction costs recall without costing any shared class its precision, so a model that leaks little and concentrates its errors inside the six mango classes, as HMLA-Net does, gains accuracy relative to macro-F1. Restricting the logits to the six shared classes (protocol P3) raises macro-F1 by 5.0 pp on average (0.2 to 28.7 pp) and reorders the leading group, lifting DenseNet-121 to the top (0.720) with HMLA-Net eleventh (0.676), and the P1 and P3 orderings correlate at $\rho = 0.82$. Protocol P2, macro-averaging over all 21 labels, equals $\frac{6}{21}$ of P1 for every model, as Section 3.7 predicts, and changes no rank. The ranking of the leading group is therefore protocol-dependent in exactly the way the replicate analysis would predict for differences of this size, and we report all three so that the reader can choose the denominator that matches their use.

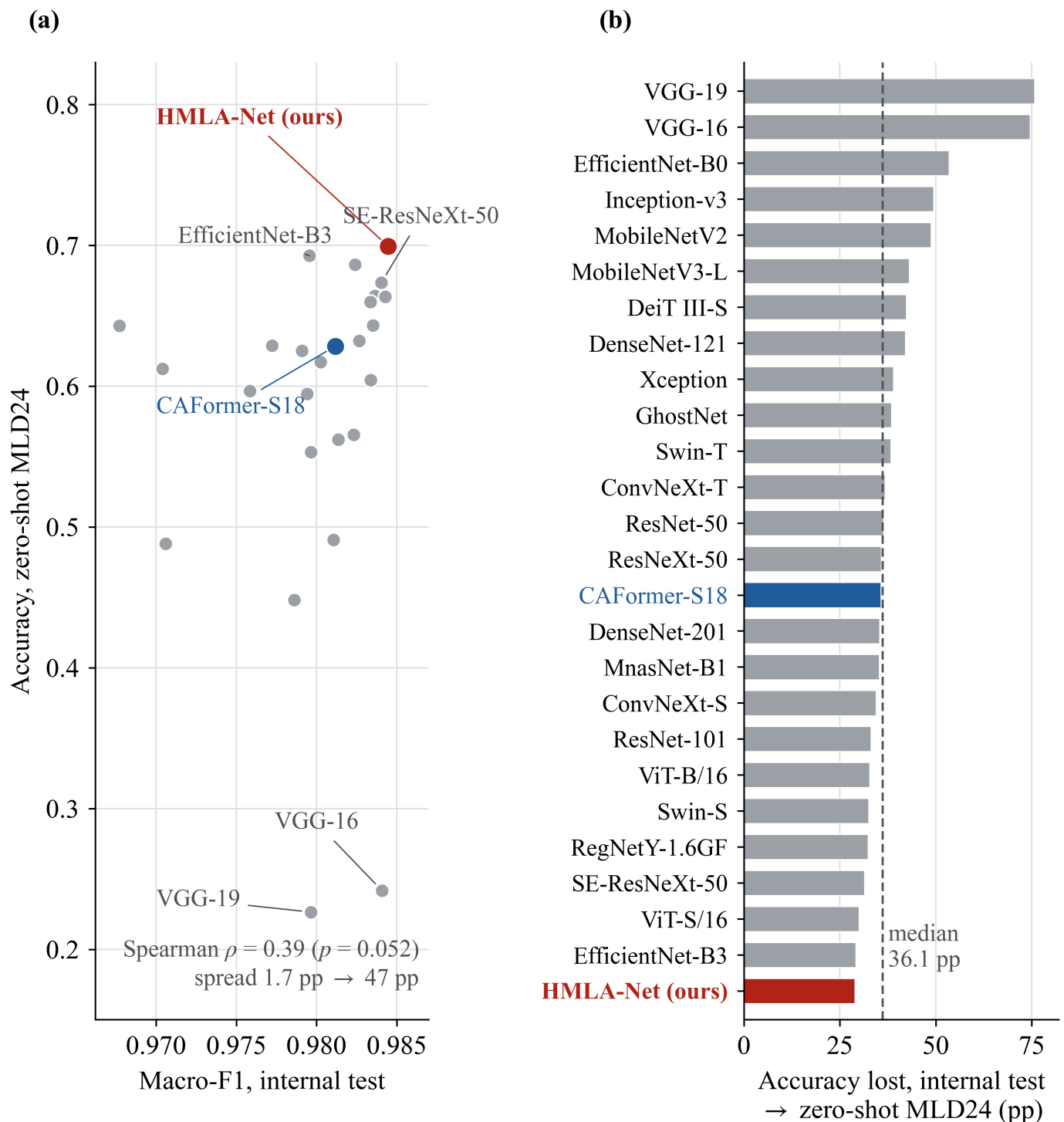


Figure 8. Transfer from the internal test partition to zero-shot MLD24 for the 26 architectures of the reference execution. **(a)** Internal macro-F1 against external accuracy. The internal axis spans 1.7 percentage points and the external axis 47, so models that are indistinguishable on the former are separated by up to 45 points on the latter. **(b)** Accuracy lost between the two partitions, sorted. HMLA-Net loses the least. The dashed line marks the median loss (36.1 pp).

4.5. Architecture or Recipe: The Matched Comparison

Table 6 reports the five backbones retrained with the complete HMLA-Net recipe. This experiment belongs to the second execution, so the table compares its rows with the HMLA-Net entry of that execution (0.6818 of external macro-F1, whereas the reference execution's entry is 0.6739). Internally, the recipe changes macro-F1 by between -0.33

and +0.33 pp, i.e., by nothing that the replicate analysis can distinguish from zero. Externally, the mean score of none of the five improved. The change is -0.13 pp for CAFormer-S18, -1.25 for Swin-S, -3.18 for EfficientNet-B3, -3.46 for ViT-S/16 and -5.11 for SE-ResNeXt-50, a mean of -2.6 pp, with none of the 20 paired comparisons surviving Holm's correction.

The controlled comparison is HMLA-Net against CAFormer-S18 under the identical recipe, which gives $+2.87$ pp of macro-F1 and $+8.1$ pp of accuracy in the benchmark entries, and $+2.05$ pp of macro-F1 with all replication-B runs pooled (nine runs against six). This is the entire measurable effect of the four architectural components, and it is inside the 3.3 pp mean replicate difference of the partition. Two further facts bound it from the other side. Recipe-matched Swin-S (0.6848) is marginally above HMLA-Net (0.6818), and the HMLA-Net replicate trained as the ablation reference of the second execution (0.6345) is below recipe-matched CAFormer-S18 (0.6531). We therefore make no claim that the architecture transfers better than its backbone. What the data support is that the complete HMLA-Net configuration belongs to the leading group of this benchmark, that its backbone under the same recipe has a lower observed mean, and that three seeds cannot resolve whether the difference is real. The leakage rate is the one quantity on which HMLA-Net is separated from every other configuration in both executions (7.4% against 21.8% for its backbone in the reference execution, 7.9% against 21.9% for the matched backbone in the second), and we note it as a property of the complete configuration without attributing it to any one component.

Table 6. Recipe-matched comparison. Five backbones were retrained with the complete HMLA-Net recipe so that the training and inference recipe is identical between them and the proposed model.

Backbone	Internal Test Macro-F1			Zero-Shot MLD24 Macro-F1			
	Generic	HMLA Recipe	Δ (pp)	Generic	HMLA Recipe	Δ (pp)	HMLA-Net – Backbone (pp)
CAFormer-S18 [†]	0.9806 ± 0.0062	0.9839 ± 0.0018	+0.33	0.6543 ± 0.0083	0.6531 ± 0.0279	−0.13	+2.87
Swin-S	0.9846 ± 0.0028	0.9813 ± 0.0030	−0.33	0.6974 ± 0.0220	0.6848 ± 0.0028	−1.25	−0.30
SE-ResNeXt-50	0.9868 ± 0.0008	0.9838 ± 0.0046	−0.30	0.6912 ± 0.0201	0.6400 ± 0.0356	−5.11	+4.17
EfficientNet-B3	0.9775 ± 0.0069	0.9795 ± 0.0046	+0.20	0.6618 ± 0.0110	0.6299 ± 0.0116	−3.18	+5.18
ViT-S/16	0.9819 ± 0.0043	0.9832 ± 0.0008	+0.13	0.6609 ± 0.0252	0.6264 ± 0.0291	−3.46	+5.54
HMLA-Net (ours)	–	0.9828	–	–	0.6818	–	–

Replication B, mean ($\pm$ s.d.) over the same three seeds and partitions. The generic column is the common recipe of Section 3.5, and the HMLA recipe column adds MixUp/CutMix, EMA weights, layer-wise learning-rate decay and V_4 flip test-time augmentation, i.e., the complete training and inference recipe of the proposed model with its architectural components (MSF, LAP, crop head, MixStyle) switched off. Δ is HMLA recipe minus generic, so a positive value means that the recipe helps. The last column is HMLA-Net minus the backbone under the identical recipe, the controlled comparison of Section 4.5. Against CAFormer-S18, it is $+2.87$ pp, and with all replication-B runs pooled (HMLA-Net 0.6601 over nine runs, CAFormer-S18 with the HMLA recipe 0.6396 over six), it is $+2.05$ pp, inside the replicate-to-replicate variation of 3.3 pp. None of the 20 paired comparisons survives Holm correction (three blocks, Section 3.8). The proposed model is set in bold; [†] denotes the bare backbone of the proposed model.

4.6. Ablation Is Uninformative In-Domain and, at Three Seeds, Not Reproducible Out-of-Domain

Table 7 and Figure 9 report every ablation variant of replication A on both partitions. On the internal partition, excluding pretraining, no removal changes macro-F1 by more than 0.29 percentage points, and eleven of the sixteen removals yield a negative estimated contribution, i.e., the variant scores above the full model. This is what one expects when the measurement is dominated by variation of the size measured in Section 4.3, although a component whose removal helps may also simply be unnecessary. On MLD24, the same removals cost up to 5.06 points and twelve of sixteen have the expected sign. The correlation between the two sets of deltas is $\rho = 0.44$ ($p = 0.08$), so an in-domain ablation table does not predict which components will survive a change of corpus.

Read on the external column, the ranking of components in this execution is ImageNet pretraining (28.8 pp, reported but omitted from the figure for scale), layer-wise learning-rate decay (5.06), MixUp/CutMix (4.82), MixStyle (4.11), label smoothing (3.27), orbit-averaged flip TTA (3.11), the MixUp cool-down (2.51), LAP (2.10), EMA (1.27), β calibration (1.24) and random erasing (1.19). Removing multi-scale fusion or the auxiliary crop head changes external macro-F1 by -0.08 and -0.13 points. Taken at face value, this ordering suggests that the training recipe carried most of the transfer gain and that MixStyle and LAP were the architectural components that mattered. The second execution does not support that reading. The three largest deltas exceed the 3.3 pp mean replicate difference of the partition, but they lie well inside its range (21 of the 78 paired replicate differences exceed five points and the largest is 20.3), and the ablation of the second execution (all 16 removals, all three seeds, provided in the repository) returns a different ordering. Against the pooled nine-run HMLA-Net reference of that execution, the same seven components measure 1.91 (layer-wise decay), 1.77 (MixUp/CutMix), -0.81 (MixStyle), 0.96 (label smoothing), 3.07 (flip TTA), 0.62 (cool-down) and -1.29 pp (LAP). Against that execution's own full-model training, which, at 0.6345, is the lowest of its three HMLA-Net trainings, every delta is 2.56 points lower still (flip TTA 0.51, layer-wise decay -0.65 , MixUp/CutMix -0.79 , MixStyle -3.37). The rank correlation between the two executions' external deltas across the 15 removals other than pretraining is $\rho = 0.34$ ($p = 0.22$) under either reference. The sign agrees for ten of the 15 with the pooled reference and for six with the full-model reference. Only pretraining (28.8 pp in this execution, 29.6 in the other) produces an effect that is large relative to replicate variation and the same in both. We therefore report Table 7 as a descriptive measurement of one execution, we attribute the transfer gain to no recipe element, and we do not infer a reliable contribution, positive or negative, for any individual component at this replication budget. The auxiliary crop head's combination coefficient β was calibrated to zero on all three seeds in both executions. The archived selection record shows that all nine grid values gave identical validation macro-F1, and the selection rule, which takes the first grid value among ties, chose zero. The combination of Equation (4) therefore changed no validation prediction that affected the score, and the calibrated inference rule reduced to the disease-head prediction. The head is at most an auxiliary training signal.

Table 7. Component ablation of replication A measured on both partitions. Removing a component changes internal macro-F1 by at most 0.29 pp once pretraining is excluded, whereas the same removals change zero-shot macro-F1 by up to 5.06 pp. The two columns of Δ are only weakly related (Spearman's $\rho = 0.44$, $p = 0.08$), and the external ordering is not reproduced by the second execution (Section 4.6).

Component Removed	Internal Test		Zero-Shot MLD24			
	Macro-F1	Δ (pp)	Macro-F1	Δ (pp)	Acc.	Δ (pp)
<i>full model</i>	0.9851 ± 0.0044	—	0.6751 ± 0.0047	—	0.6937	—
ImageNet pretraining	0.9236 ± 0.0044	+6.15	0.3873 ± 0.0196	+28.78	0.3931	+30.07
layer-wise learning-rate decay	0.9861 ± 0.0033	-0.10	0.6245 ± 0.0581	+5.06	0.6201	+7.36
MixUp/CutMix	0.9839 ± 0.0065	+0.12	0.6269 ± 0.0408	+4.82	0.6535	+4.03
MixStyle	0.9859 ± 0.0034	-0.07	0.6339 ± 0.0356	+4.11	0.6127	+8.10
label smoothing	0.9868 ± 0.0058	-0.17	0.6423 ± 0.0452	+3.27	0.6724	+2.13
flip test-time augmentation	0.9859 ± 0.0037	-0.08	0.6440 ± 0.0316	+3.11	0.6537	+4.01
MixUp cool-down [‡]	0.9846 ± 0.0051	+0.05	0.6500 ± 0.0287	+2.51	0.6561	+3.76
LAP (lesion-aware pooling)	0.9852 ± 0.0033	-0.01	0.6541 ± 0.0275	+2.10	0.6556	+3.81
EMA weights	0.9840 ± 0.0052	+0.11	0.6624 ± 0.0581	+1.27	0.6874	+0.63
β calibration	0.9880 ± 0.0051	-0.29	0.6626 ± 0.0143	+1.24	0.6930	+0.08
random erasing	0.9851 ± 0.0064	+0.00	0.6632 ± 0.0181	+1.19	0.6615	+3.23
two-phase schedule	0.9862 ± 0.0055	-0.11	0.6718 ± 0.0139	+0.33	0.7003	-0.65
MSF (multi-scale fusion)	0.9875 ± 0.0021	-0.23	0.6759 ± 0.0269	-0.08	0.6841	+0.97

Table 7. Cont.

Component Removed	Internal Test		Zero-Shot MLD24			
	Macro-F1	Δ (pp)	Macro-F1	Δ (pp)	Acc.	Δ (pp)
auxiliary crop head	0.9864 ± 0.0014	−0.13	0.6764 ± 0.0214	−0.13	0.6862	+0.76
LAP gating and multi-head attention	0.9853 ± 0.0023	−0.01	0.6789 ± 0.0147	−0.38	0.6894	+0.43
class-balanced sampler	0.9863 ± 0.0043	−0.12	0.6835 ± 0.0261	−0.84	0.7028	−0.90

Every variant is trained from scratch with the same three seeds and the same partitions as the full model. Δ is the metric of the full model minus the metric of the variant, so a *positive* Δ means the removed component contributes. Architectural components are set in bold; the remainder belong to the training recipe. No Δ on the internal partition survives the Holm–Bonferroni-corrected Wilcoxon test, and with three paired blocks, the smallest attainable two-sided p is 0.25, so the internal column is reported as a descriptive measurement rather than a test. [‡] The MixUp cool-down variant completed two of the three seeds on the internal partition and all three on MLD24, so its internal value is not fully paired and is excluded from the paired statistics. In the second execution of the benchmark (Section 4.3), all 16 removals completed all three seeds. That execution’s ablation is provided in the repository and its external ordering of the components correlates with this one at Spearman’s $\rho = 0.34$ under either reference definition of Section 4.6.

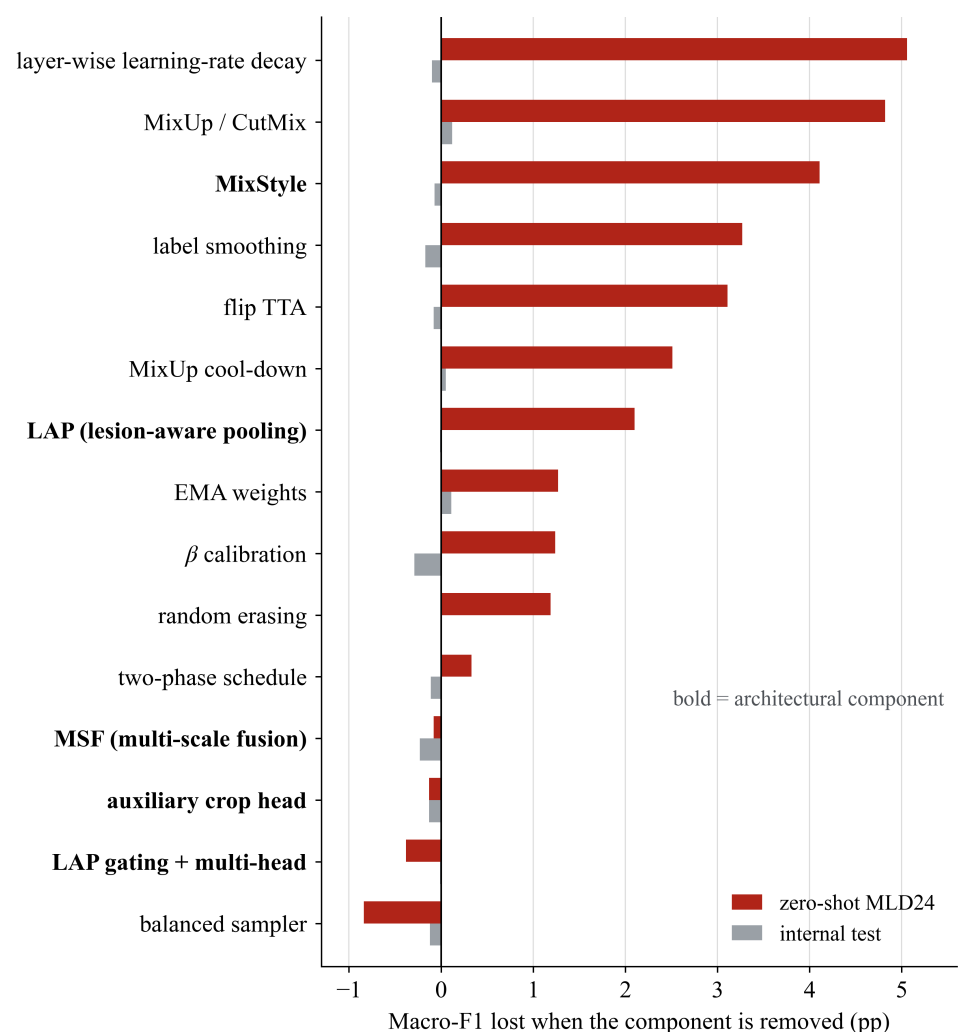


Figure 9. Ablation of replication A measured on both partitions. Each bar pair represents one component being removed from the full model and retrained, and positive means the component contributes. Internal deltas (grey) are indistinguishable from zero for every component, and negative for the majority. External deltas (red) reach 5 percentage points, which exceeds the mean replicate-to-replicate difference of the partition (3.3 pp, Section 4.3) but lies inside its range, and their ordering is not reproduced by the second execution ($\rho = 0.34$). Architectural components are in bold. The ImageNet pretraining variant (+6.15 internal, +28.78 external) is omitted for scale and reported in Table 7.

4.7. One Class Does Not Survive the Change of Corpus

Table 8 and Figure 10 decompose the transfer result of HMLA-Net by class. On the internal test partition, all six shared mango classes reach an F1 of 1.000 on every seed. On MLD24, four remain usable (die back 0.985, healthy 0.946, anthracnose 0.872, gall midge 0.664), powdery mildew degrades to 0.567, and sooty mould collapses to 0.010, with a recall of 0.011, 0.000 and 0.005 on the three seeds (0.015, 0.000 and 0.000 in the second execution), against a mean F1 of 0.807 for the other five classes. The destination of the missed images is concentrated, with 64% of the MLD24 sooty-mould leaves assigned to powdery mildew and 13% to gall midge, and 7% going to the jackfruit sooty-mould class. In the second execution, the shares are 65, 21 and 4%.

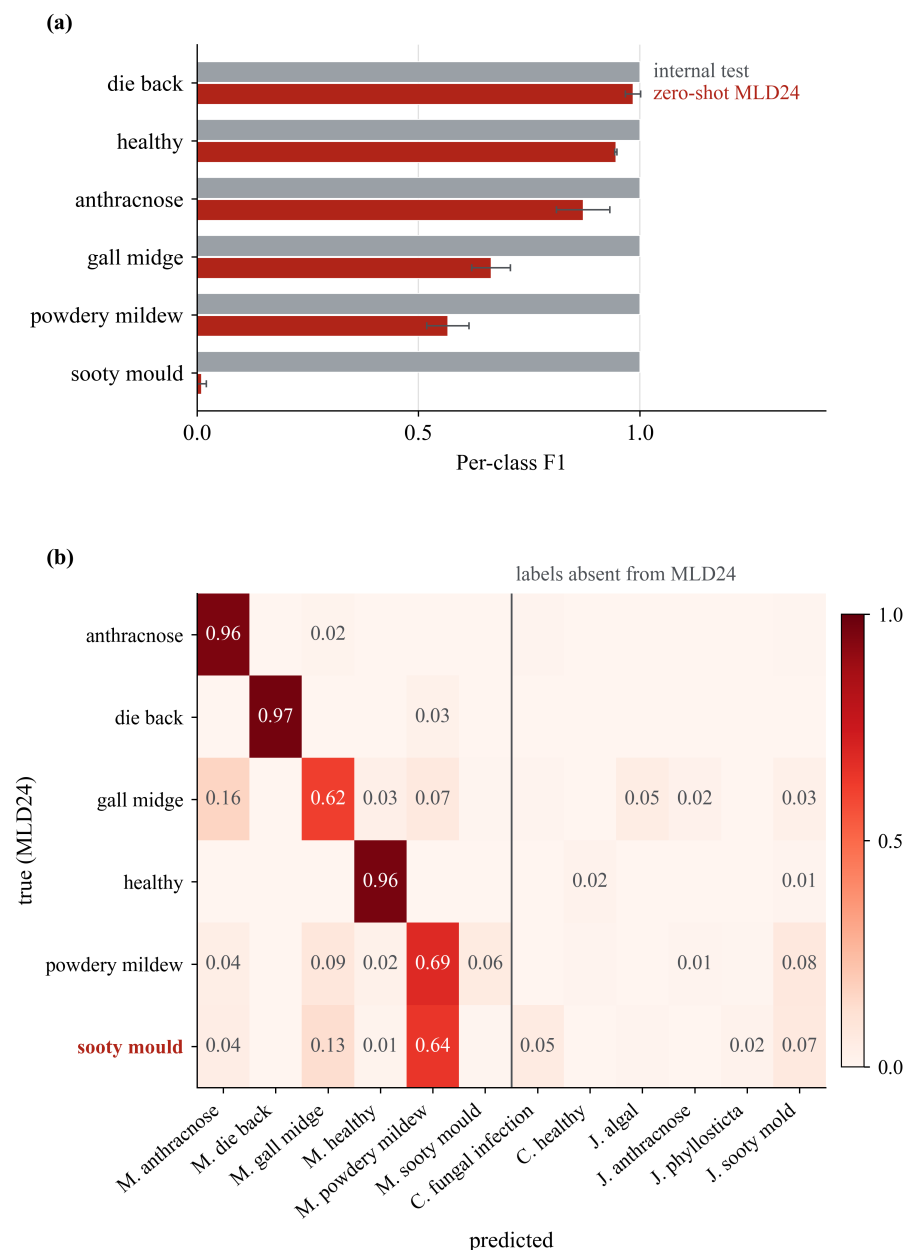


Figure 10. Per-class behaviour of HMLA-Net on the six mango classes shared by the two corpora (reference execution). **(a)** F1 in-domain (grey) and zero-shot (red, mean $\pm$ s.d. over three seeds). **(b)** Row-normalised confusion matrix on MLD24, averaged over the three seeds and computed directly from the archived per-image predictions. Columns to the right of the rule are labels absent from MLD24 that nevertheless receive predictions, which the unchanged 21-way head makes possible. Sooty mould is routed predominantly to powdery mildew.

This single class is why HMLA-Net leads on accuracy and is fourth on macro-F1. The six classes have equal support, so a class at zero recall costs both metrics the same one-sixth of its own weight. The difference arises on the other side of the confusion matrix. Accuracy counts only how many images are wrong, wherever they land. Macro-F1 also charges the class that receives them. The 514 sooty-mould leaves routed to powdery mildew halve that class's precision (0.485 in Table 8) and pull its F1 down to 0.567 although its own recall is 0.687, so one collapsed class damages two of the six F1 terms while it damages the accuracy sum once. A model that spreads its errors thinly across classes, including into labels absent from MLD24 where they cost recall but no shared class's precision, loses less macro-F1 for the same accuracy. That is the trade the models above HMLA-Net on macro-F1 make, and Table 5 shows it in their higher leakage rates.

Table 8. Per-class behaviour of HMLA-Net on the six mango classes that both corpora share.

Mango Class	Internal Test				Zero-Shot MLD24			
	<i>n</i>	Prec.	Rec.	F1	<i>n</i>	Prec.	Rec.	F1
die back	74	1.000	1.000	1.000	800	1.000	0.970	0.985 ± 0.017
healthy	75	1.000	1.000	1.000	800	0.932	0.962	0.946 ± 0.002
anthracnose	73	1.000	1.000	1.000	800	0.810	0.957	0.872 ± 0.060
gall midge	75	1.000	1.000	1.000	800	0.727	0.616	0.664 ± 0.043
powdery mildew	75	1.000	1.000	1.000	800	0.485	0.687	0.567 ± 0.048
sooty mould	32	1.000	1.000	1.000	800	0.089	0.005	0.010 ± 0.011

Mean over three seeds; internal support is the mean test-partition count. The six classes are those MLD24 shares with our label space. *Sooty mould* is classified perfectly in-domain and is essentially unrecognised out-of-domain (recall 0.011/0.000/0.005 in the three seeds) while the other five average 0.838. Section 4.7 shows that a collapse of one class to exactly zero recall is shared by 19 of the 26 architectures, which points to a difference in what the two corpora label rather than to a property of any one model.

The collapse is not specific to the model.

Scanning every run of the 26 architectures, 19 produce at least one MLD24 class at exactly zero recall in at least one seed and five do so in every seed (20 and two in the second execution), and in every such instance, the class is mango sooty mould. The seven architectures that keep every class above zero (DenseNet-121, EfficientNet-B0, GhostNet, MnasNet-B1, ResNet-101, Swin-T and ViT-B/16) do not thereby transfer better, since their macro-F1 ranges from 0.531 to 0.662. The failure is therefore concentrated in one class across most architectures, and the scan identifies the class, not the cause.

What the images show.

Figure 11 places six internal and six MLD24 images labelled sooty mould side by side, and quantitative descriptors computed on 150 images per corpus and class make the visible difference measurable. The MLD24 images are sharper (Laplacian variance 0.0067 against 0.0010, Cohen's $d = 4.1$), darker ($d = -1.4$), more saturated ($d = 1.1$) and carry a larger fraction of dark pixels (0.156 against 0.114, $d = 0.7$). Sharpness, however, differs by $d = 1.5$ to 4.9 for every shared class, and brightness by $d = -0.8$ to -2.3 , so the corpus-wide acquisition shift does not by itself identify the cause of a class-specific collapse. What is specific to sooty mould is the appearance of the symptom. The internal images show a diffuse dark film over much of the lamina, whereas the MLD24 images show discrete pale patches on an otherwise green leaf, which is the appearance the model associates with powdery mildew, and indeed where it sends them. In the embedding space of a pretrained network, only 40% of the MLD24 sooty-mould images have an internal nearest neighbour of the same class, and the class centroids of the two corpora have cosine similarity 0.88, lower than for the other five classes. These observations are consistent with three non-exclusive explanations. The two curation teams may have drawn the boundary of the label at different points on a continuum of symptoms (a label-definition difference),

they may have photographed different stages or severities of the same condition, and the corpus-wide acquisition shift may interact more strongly with a surface film than with a discrete lesion. Sooty mould is a secondary symptom, a superficial fungal growth on the honeydew deposited by sap-feeding insects such as aphids, leafhoppers, mealybugs and soft scales [76], so the second explanation is biologically natural, because what is visible depends on how much honeydew and mould have accumulated, and neither corpus records the insect damage that precedes it. Our experiments cannot separate these explanations, and the present paper therefore states label-definition mismatch as a hypothesis, not as a finding. Adjudicating it would require an expert assessment of the two corpora's images, which is outside the scope of this work.



Figure 11. Images labelled “mango sooty mould” in the internal corpus (top row, training material) and in MLD24 (bottom row, zero-shot evaluation), drawn at random. The internal images show a diffuse dark coating over the lamina, whereas the MLD24 images show discrete pale patches on green tissue, which is the appearance HMLA-Net associates with powdery mildew, the class to which about two-thirds of them are assigned. The panel is presented as evidence for a hypothesis, not as an adjudication of the labels.

4.8. Computational Cost and the Role of Model Size

HMLA-Net has 23.9 M parameters, 3.97 GFLOPs and a 91.2 MB float32 footprint. Six of the eight strongest internal models are larger; the eight have a median of 38.2 M parameters and a maximum of 134.3 M (VGG-16). The bare backbone with its default classification head and 21 outputs is 24.3 M, of which the head, a two-layer perceptron with a 2048-unit hidden layer, accounts for 1.10 M. HMLA-Net discards that head and adds 0.66 M (multi-scale fusion 0.32 M, lesion-aware pooling 0.33 M, the disease and crop classifiers with their normalisation 0.01 M), so the complete model is 0.44 M smaller than the backbone it is built on. Measured single-pass inference latency is 11.1 ms at a batch size of one on the RTX A4000. The four-view orbit average used throughout raises this to 44.4 ms, or 22.5 images per second, the slowest in the set on a per-image basis, and Section 4.9 reports what it buys. Mean training time was 864 s per seed.

Figure 12 places the 26 architectures on the parameter–transfer plane. Zero-shot accuracy shows no evidence of a monotonic association with parameter count (Spearman’s $\rho = 0.13$, $p = 0.53$, and a linear fit in $\log_{10}$ parameters explains 5% of the variance), but a quadratic fit explains 56% (F -test against the linear fit, $p < 10^{-4}$), is concave, and has its vertex at 13.9 M parameters (bootstrap 95% interval 10.9 to 54.0 M, and 95% of bootstrap resamples return a concave fit with an interior vertex). In parameter bins, the six models below 10 M average an accuracy of 0.547 ± 0.067 , the 17 between 10 and 50 M average 0.631 ± 0.052 , and the three above 50 M average 0.376 ± 0.246 (Kruskal–Wallis $p = 0.021$).

The four smallest models lose between 35 and 54 accuracy points under transfer and the two largest, VGG-16 and VGG-19 at 134 to 140 M, lose 75 to 76. On the internal partition, the relationship is a weak monotone one instead (Spearman's $\rho = 0.45$, $p = 0.023$, and the quadratic term adds nothing, $p = 0.32$). The concave shape rests on the two VGG networks. Without them, the quadratic fit explains 30% of the variance and has no interior vertex, and zero-shot accuracy is weakly and positively associated with parameter count ($\rho = 0.44$, $p = 0.033$), as it is on the internal partition. Capacity is thus neither sufficient nor necessary for transfer. The apparent optimum near 14 M is a description of this set of 26 architectures, dominated by the failure of its two largest members, and not a design rule.

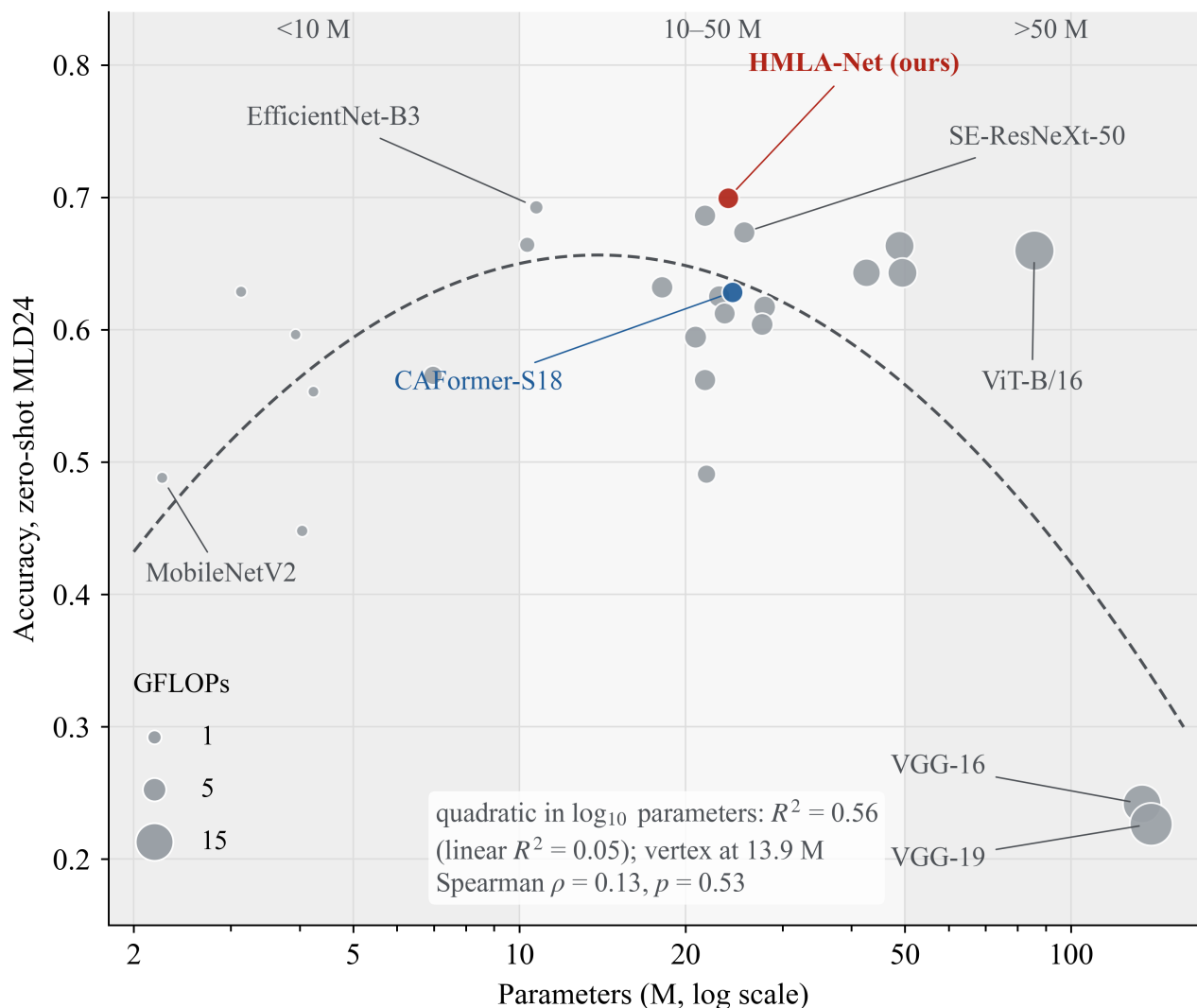


Figure 12. Parameter count (logarithmic) against zero-shot accuracy for the 26 architectures of the reference execution. Marker area is proportional to GFLOPs. The dashed curve is the quadratic fit in $\log_{10}$ parameters ($R^2 = 0.56$, vertex at 13.9 M), and the shaded bands are the three size bins discussed in the text. The concave shape depends on the two VGG networks (Section 4.8).

4.9. The Symmetry Constraint: Cost and Benefit of Orbit Averaging

Table 9 evaluates one HMLA-Net checkpoint of the second execution under three inference groups, namely the trivial group (a single forward pass), the Klein four-group V_4 of Equation (6), and the full dihedral group D_4 of flips and quarter-turn rotations. Two things are established exactly. The orbit-averaged posterior is invariant under every element of the group to floating-point precision (largest deviation 6×10^{-8} for V_4 and 1.2×10^{-7} for D_4), whereas the single-view network, which was trained with the same

flips as augmentation, moves its posterior by up to 0.014 under a flip. Training-time augmentation makes the network approximately equivariant and inference-time orbit averaging makes it exactly invariant, and the two should not be described as the same operation. The measured cost in this experiment is a factor of 3.9 in latency for V_4 and 10 for D_4 . The benefit is a single-run measurement, +0.5 pp of external macro-F1 and +0.8 of accuracy for V_4 and +1.2 and +2.1 for D_4 , with a change of -0.2 pp internally. The three-seed ablation puts the removal of V_4 averaging at +3.1 pp of external macro-F1 in replication A and, against the pooled reference of replication B, also at +3.1, but at +0.5 against that execution's own full-model training, so the agreement between executions is reference-dependent (Section 4.6) and the effect is inside replicate variation either way. We report these as the honest size of the effect. Imposing an invariance the data possess is cheap to state, and exact to verify. It yielded a gain of a few points under shift in the evaluations reported here.

Table 9. Orbit averaging at inference: What the symmetry group costs and buys.

Group G	$ G $	Int. F1	Ext. Acc.	Ext. F1	ms	Img/s	Orbit Dev.	Raw Dev.
trivial group $\{e\}$ (no TTA)	1	0.9887	0.6975	0.6794	9.4	106.6	—	—
Klein four-group V_4 (flips)	4	0.9865	0.7054	0.6847	36.5	27.4	6.0×10^{-8}	0.014
dihedral group D_4 (flips and rotations)	8	0.9869	0.7188	0.6910	93.9	10.7	1.2×10^{-7}	0.017

One checkpoint of HMLA-Net (replication B, seed 1234) evaluated under three inference groups, with latency at a batch size of one on the RTX A4000. The orbit dev. column is the largest absolute change in the orbit-averaged posterior when the input is transformed by an element of G (floating-point noise only, i.e., exact invariance), and the raw dev. column is the same quantity for the single-view network, which is only approximately equivariant after training-time flip augmentation. The differences between the three rows (+0.53 and +1.17 pp of external macro-F1 for V_4 and D_4 against no TTA) are single-run measurements and lie within the replicate-to-replicate variation of Section 4.3. The latency ratios are measured values for this system and batch size.

4.10. Explainability

Figure 13 shows Grad-CAM maps computed at the MSF output of a checkpoint of the second execution. Activation concentrates on lesion tissue rather than on background or leaf margin, and the 14×14 resolution afforded by the fusion stage resolves individual spots that a stride-32 map would merge. Figure 14 places the LAP attention map of Equation (2), averaged over the four heads, beside the Grad-CAM map for seven of the same images. The two maps are computed by different mechanisms, one from gradients of the class score and one from the pooling weights the network learned, and they agree on where the evidence is. Where they differ, the attention map is the more selective of the two, concentrating on individual spots (jackfruit phyllosticta and leaf spot) where Grad-CAM spreads over the lamina. We report both as qualitative observations. We did not compute deletion or insertion curves, and pointing-game scores and intersection-over-union against lesion masks would require lesion annotations, which the corpus does not carry. We therefore make no quantitative claim about explanation faithfulness, and the maps shown, computed for correctly classified internal test images, do not explain the out-of-domain failures. Section 2 notes that no reviewed study in this literature does either, which is context, not justification.

Figure 15 shows the pooled 512-dimensional descriptors under t-SNE for the internal test partition. Every class forms a compact, well-separated cluster, which is the in-domain saturation of Section 4.1 seen from the feature side. Clusters of the same crop tend to be neighbours, but t-SNE preserves neither global distances nor between-cluster geometry, so we draw no conclusion from the arrangement beyond the separation itself.

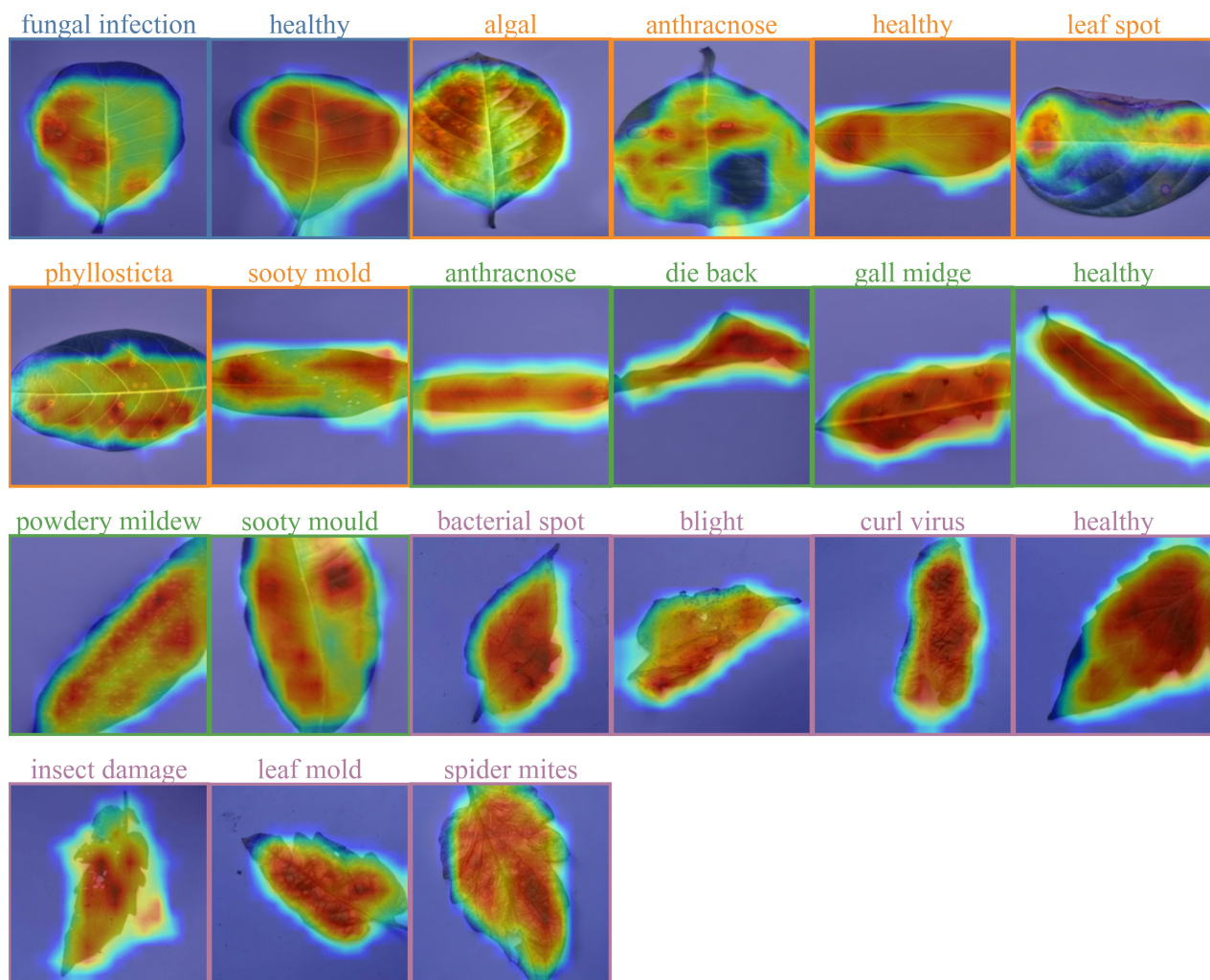


Figure 13. Grad-CAM at the multi-scale fusion output (14×14) for one representative test image of each class. Every image shown is correctly classified. Warm regions denote evidence for the predicted class, and frame colour encodes the crop (blue for cashew, orange for jackfruit, green for mango and purple for tomato).

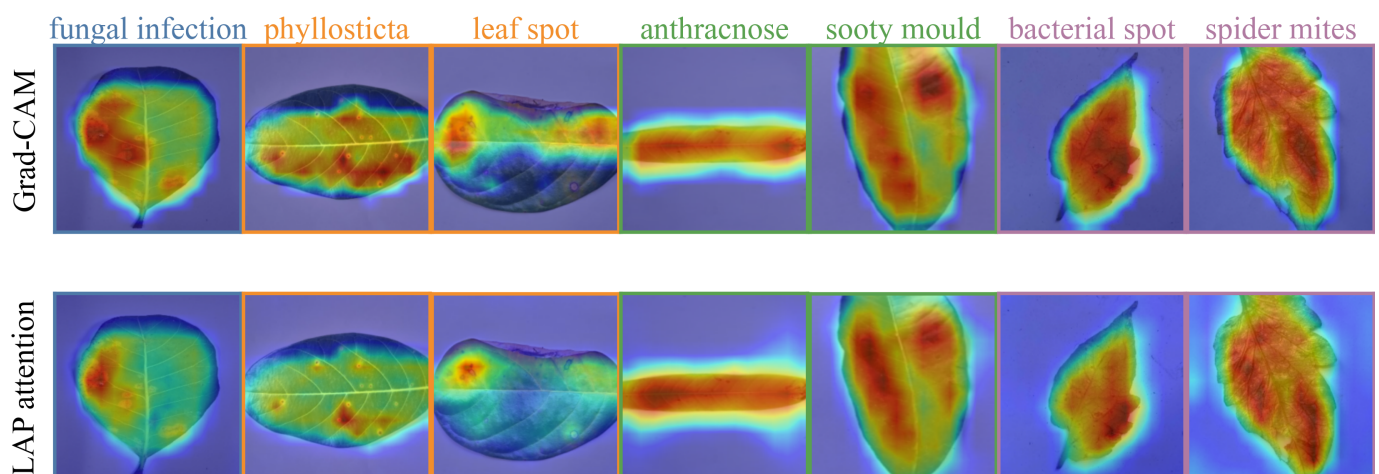


Figure 14. Grad-CAM (top) and the gradient-free LAP attention map (bottom, four heads averaged) for seven of the images of Figure 13. The two maps are derived from different mechanisms and largely agree. The attention map is the more selective where lesions are small and scattered.



Figure 15. t-SNE of the 512-dimensional pooled descriptors on the internal test partition (seed 1234), coloured by crop and labelled by disease. All 21 classes form compact, separated clusters. The arrangement of clusters relative to one another is not meaningful in a t-SNE embedding.

5. Discussion

5.1. What a Saturated Benchmark Can and Cannot Support

The most consequential numbers in this paper are 0.34 and 0.69, namely the mean difference in internal macro-F1 between two trainings of the same configuration on the same seed and partition, and the rank correlation between the two orderings that those trainings produce. Every in-domain comparison in Section 4.1 is smaller than a few multiples of the first, and the second says that a leaderboard built on this partition is only partly reproducible even by its own authors, on their own machine, with their own code. A benchmark in this state cannot order its leading methods, and a paper claiming that a new architecture leads such a benchmark by tenths of a point may be reporting a difference that would not survive re-execution. We hold our own claim to the same standard. In one

execution, HMLA-Net had the highest internal macro-F1 of 26 models and in the other, it was eighth, and neither fact is informative.

This has a direct implication for how the present literature should be read. Of the hybrid models surveyed in Section 2, the gains of fusion over the stronger single branch computed from [17] are 0.22, 1.46 and 0.55 percentage points, with an attention module adding a further 0.17 to 0.28. Those magnitudes are in the range of the replicate variation we measure, and none of the studies reports replicate-to-replicate variation. We do not claim their conclusions are wrong. We claim that the evidence offered is not sufficient to distinguish them from variation, and that the remedy is affordable. It consists of re-executing the benchmark, or at least a representative subset of it, and printing the distribution of differences rather than a single number. A single duplicated configuration is not enough. The two trainings of HMLA-Net within the reference execution, as benchmark entry and as ablation reference, differ by only 0.06 pp of the three-seed mean, whereas the same quantity over the 26 configurations averages 0.19 pp between executions (0.17 pp for HMLA-Net itself, and 0.34 pp per run), so a single number is not a generalisable estimate of replicate variation, which is itself an argument for reporting the distribution rather than the single number.

We also separate two statements that are easy to conflate. The empirical statement is that the spread between models is comparable to the spread between replicates of one model. That is a measurement and it stands on its own. The statistical statement is that three paired blocks cannot reach significance under Holm's correction and that the Nemenyi critical difference covers the rank axis. That is a property of the design, and it is not evidence that the models are equivalent. The first supports the title of this paper, and the second only explains why the conventional tests could not have contradicted it.

5.2. *Why the External Partition Carries the Information*

The same checkpoints that are indistinguishable in-domain span 47 accuracy points out-of-domain, and that ordering replicates ($\rho = 0.92$ for accuracy, 0.91 for macro-F1) although the individual numbers vary ten times more between executions than the internal ones do. The ratio of spread to replicate variation is what matters, and it is about five internally and above ten externally. The external partition is therefore informative at the level of model families. It separates the VGG networks, which lose 75 to 76 accuracy points, and the smallest efficiency-oriented networks, which lose 35 to 54, from the mid-sized convolutional, transformer and hybrid networks, which lose 29 to 43, and it does so reproducibly. It is not informative, at three seeds, about the order within the leading group, whose five members lie inside one replicate difference of one another and change places between protocols P1 and P3.

Two mechanisms are visible in our data. First, in-domain, the corpus contains cues that are perfectly predictive and perfectly non-transferable. The background tint that separates tomato from the other three crops in Figure 3 is the clearest, and a model that learns background is indistinguishable from a model that learns lesions until the background changes. Second, the transfer scores of the leading group are governed by where a model puts its errors as much as by how many it makes, that is, by leakage into the 15 absent labels, which costs recall alone, against confusion inside the six shared classes, which costs precision as well. HMLA-Net has the lowest leakage of the 31 configurations and concentrates its errors inside the mango block, which is why it leads on accuracy and not on macro-F1. Whether that behaviour is a consequence of any of its components is a question the ablation cannot answer at this budget.

The recipe-matched experiment settles a question that a single-execution ablation leaves open. The four recipe elements that the baselines had not received do not, when

given to them, improve their transfer, and the margin of HMLA-Net over its own backbone under the identical recipe is 2 to 3 points of macro-F1, inside replicate variation. The single-execution ablation suggested that the transferable part of a leaf disease classifier was mostly in how it was trained. The second execution and the matched baselines show that this too was a reading of variation. What remains is a smaller and more defensible statement. The complete HMLA-Net configuration is a member of the leading group of this benchmark, its backbone under the same recipe is not, and the attribution of that difference to architecture, recipe or chance requires more replicates than three seeds provide.

5.3. The Magnitude of the Degradation

HMLA-Net loses 28.9 accuracy points moving from the internal test partition to MLD24, and the median configuration loses 36.1. Placing this against the literature requires care, because the studies that measure transfer differ in how far they move. Sun [4] reports decreases in macro-F1 for a PlantVillage-trained ResNet from 0.987 to 0.012 and 0.130 on two other tomato corpora, and Krishna et al. [5] report accuracies 14 to 24 points lower when models trained on web-sourced images are tested on PlantDoc; both involve a change of image source, and their target corpora include field or web imagery. Our shift is milder in that both of our corpora are studio-style acquisitions of detached leaves from the same country, differing in research group, capture session, camera and delivered resolution, and the duplicate audit of Section 3.1 found no exact duplicates and no evidence of substantial overlap between them. Losing 29 to 36 points across so mild a shift is the point. It suggests that the in-domain figures reported throughout this literature are not merely optimistic about field deployment but optimistic about any change of provenance, including one as small as a different phone camera. We add the caveat that this has been measured for one crop. The source corpus contains four, MLD24 contains mango, and nothing here shows that the same ordering, or the same magnitude, holds for tomato, jackfruit or cashew.

5.4. The Sooty Mould Finding

That one class survives training perfectly and transfer not at all is, we think, the most actionable observation here for corpus builders. Mango sooty mould is the smallest class in our corpus (209 images) and is classified without error in-domain. On MLD24, it is recognised in about half a percent of 800 images and is routed predominantly to powdery mildew, in 19 of 26 architectures in the reference execution and 20 in the second. Section 4.7 shows what differs between the two sets of images, quantitatively and visually, and offers three explanations that the present experiments cannot separate, namely a difference in what the two curation teams call sooty mould, a difference in the stage or severity they photographed, and an interaction between a corpus-wide acquisition shift and a symptom that is a surface film rather than a lesion. It is tempting to conclude that the two corpora do not label visually equivalent material under the same name and that no architecture could repair such a mismatch. The first is a hypothesis, and we state it as one. The second follows only if the first is true. Because sooty mould is a secondary symptom of honeydew-producing insects, an adjudication would need the insect damage that precedes it to be recorded, which neither corpus does.

The practical consequence is that aggregate cross-corpus numbers can be dominated by a single class. Had we reported only macro-F1, the leading group would have been ordered by how each model happens to distribute one class's errors, and the paper's comparison would have turned on a quantity that measures the agreement between two annotation protocols as much as any property of a classifier. Reporting accuracy, macro-F1 under more than one protocol, the leakage rate, the geometric mean of per-class recalls and

the full per-class breakdown is what makes the distinction visible, and we would encourage the practice.

5.5. Limitations

Eight limitations bound what this paper establishes. One of them, near-duplicates in the internal partition, is measured rather than open.

Statistical power: Three seed blocks make seed-level Wilcoxon significance unreachable by construction, and surviving Holm's correction over the 25 comparisons of Table 3 would require ten blocks, a compute multiplier of roughly $3.3\times$, and that is only the point at which rejection becomes arithmetically possible. Every seed-level test we report is descriptive, and the per-image McNemar test, whose power depends on the 5 to 11 discordant pairs rather than on the 1077 test images, separates nothing either.

Replicate variation is estimated, not bounded: Two executions give one paired difference per configuration and seed. The distribution over 78 pairs is far better than a single duplicated run, but it estimates the typical size of replicate variation and not its lower bound, and the two executions were three weeks apart on one machine. A different machine or driver would add variation we have not measured.

Uncertain contributions of individual components: Under the identical recipe, the margin of HMLA-Net over its backbone is inside replicate variation, and the single-component ablation attributes nothing but pretraining on either partition. The architectural additions are retained for their design rationale and, in the case of MSF, for the 14×14 attribution grid, not for any measured accuracy contribution.

One external corpus, one crop, six classes: Transfer is measured on mango only, because that is the overlap MLD24 offers. Whether the ordering of Table 5 holds for tomato, jackfruit or cashew is untested, and the title's "multi-crop" describes the training corpus, not the transfer evidence.

Near-duplicates within the internal corpus: The partition is not near-duplicate aware, and 17 to 22% of the test images have a perceptual-hash neighbour in training. Re-scoring every run without those images changes each configuration's macro-F1 by -0.11 to $+0.43$ pp and leaves the ordering essentially unchanged (Section 3.2), so the saturation does not depend on them. A near-duplicate-aware partition would nevertheless be the cleaner design, and we recommend it for future corpora.

The sooty mould hypothesis is not adjudicated: The visual and descriptor evidence of Section 4.7 is consistent with a label-definition difference and with two other explanations, and the present experiments do not separate them.

No comparison against published architectures: Our 25 baselines are generic backbones, not re-implementations of FAMNet [17], LeafFusionNet [21] or the mixture-of-experts router [23]. Table 1 is therefore contextual rather than a head-to-head comparison, and we label it as such.

Qualitative explainability only: No deletion or insertion curves were computed, and pointing-game scores or IoU against lesion masks would require lesion annotations that the corpus lacks (Section 4.10). Calibration is also mediocre in absolute terms. Expected calibration error is 0.079 in-domain and 0.083 out-of-domain, so predicted confidences should not be used as probabilities without post hoc temperature scaling.

6. Conclusions

We trained 26 architectures on a four-crop, 21-class leaf disease corpus under one protocol with shared partitions and three seeds, executed the whole benchmark twice, and found the held-out partition saturated in a precise sense. The 26 architectures span 1.68 percentage points of macro-F1, the same configuration differs by 0.34 points on average

and 1.06 at most between executions, and the ordering of the models replicates only partly ($\rho = 0.69$). Transferring the same checkpoints, without adaptation, to an independently collected mango corpus spread them over 47 accuracy points with an ordering that does replicate ($\rho = 0.92$), separating the families of models reproducibly while leaving the leading group unordered.

HMLA-Net, which adds multi-scale fusion, gated multi-head lesion-aware attentive pooling, an auxiliary crop head with a validation-calibrated combination coefficient, and MixStyle to a CAFormer-S18 backbone, and which averages its posterior over the Klein four-group of flips at inference, attains in the reference execution the highest internal macro-F1 (0.9845) and the highest zero-shot accuracy (0.699) of the 26 architectures, with the smallest internal-to-external degradation (28.9 against a median of 36.1 points) and the lowest rate of predictions leaking into labels the target corpus does not contain. The second execution reproduces its zero-shot rank and its degradation but not its internal rank, which we therefore report as a fact about one execution. Five backbones retrained with its complete recipe show that its margin over its own backbone under identical conditions is inside replicate variation, and in a 16-removal ablation measured on both partitions, only the large effect of ImageNet pretraining remained consistent across the two executions. We therefore claim membership of the leading zero-shot group for the complete configuration and nothing for the architecture alone. One class, mango sooty mould, is perfect in-domain and essentially unrecognised out-of-domain in 19 of the 26 architectures, and we present a label-definition difference between the corpora as one of three hypotheses for it.

Three recommendations follow for the field. The first is to replicate the benchmark, not one configuration, because a matched repeat costs the same training budget as the original (23.5 GPU hours here) and turns an unfalsifiable leaderboard into a measured one. The second is to evaluate out of domain and to report the replicate variation of that partition beside its scores, because a single execution's seed-level error bars understated it here by a factor of about seven. The third is to report per-class transfer under more than one scoring protocol, because a single class and a single choice of denominator can reorder an aggregate cross-corpus metric.

Several directions for future work follow directly from the limitations above. The first is a second external corpus that covers tomato, jackfruit and cashew, ideally acquired in the field, so that the transfer ordering reported here for mango can be tested on the other three crops and under a laboratory-to-field shift. The second is a larger replication budget, with ten or more seed blocks for the component ablation and a third end-to-end execution of the benchmark, which would turn the descriptive tests of this paper into inferential ones and bound replicate variation instead of estimating it. The third is the evaluation of the leading published hybrid architectures under this protocol, with shared partitions, identical recipes and paired repeats, which is the comparison the field currently lacks. The fourth is an expert adjudication of the sooty mould labels of the two corpora by plant pathologists, together with a record of the insect damage that precedes the symptom, so that the label-definition hypothesis can be tested rather than stated. Finally, lesion-level annotations would allow the attribution maps to be scored quantitatively with deletion and insertion curves and pointing-game metrics, and post hoc calibration and near-duplicate-aware partitioning would address the two remaining weaknesses of the training protocol. Because the code, partitions and per-image predictions of both executions are public, each of these extensions can be read against the replicate variation reported here rather than against a single number.

Author Contributions: Conceptualisation, S.K. and D.A.; methodology, S.K. and M.B.; software, M.B. and D.A.; validation, M.B.; formal analysis, M.B. and D.A.; investigation, S.K.; data curation, S.K.; writing—original draft preparation, S.K.; writing—review and editing, M.B. and D.A.; visualisation, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Both corpora are publicly available under CC BY 4.0. The internal corpus is the Multi-Crop Leaf Disease Dataset for Deep Learning-Based Classification and Explainable AI, <https://data.mendeley.com/datasets/3xd9n7jpc8/1> (accessed on 2 August 2026; <https://doi.org/10.17632/3xd9n7jpc8.1>) [37]. The external corpus is MLD24, <https://data.mendeley.com/datasets/6dvpwym2m2/1> (accessed on 2 August 2026; <https://doi.org/10.17632/6dvpwym2m2.1>) [38]. The training and evaluation code, the configuration of every run, the recipe matrix, the partition files for the three seeds with their SHA-256 fingerprints, the archived per-run metrics and per-image predictions of both replications, the duplicate-audit and class-level analysis outputs, and the scripts that generate every table, figure and statistical result in this paper from those outputs are publicly available at <https://github.com/songulkarakus/HMLA-Net> (version 1.1.2; accessed on 6 September 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the ROC Curve
CBAM	Convolutional Block Attention Module
CI	Confidence Interval
CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
ECE	Expected Calibration Error
EMA	Exponential Moving Average
GAN	Generative Adversarial Network
GELU	Gaussian Error Linear Unit
GFLOPs	Giga Floating-Point Operations
Grad-CAM	Gradient-weighted Class Activation Mapping
HMLA-Net	Hierarchical Multi-Scale Lesion-Attention Network
LAP	Lesion-Aware Attentive Pooling
LIME	Local Interpretable Model-agnostic Explanations
LR	Learning Rate
LSTM	Long Short-Term Memory
mAP	mean Average Precision
MCC	Matthews Correlation Coefficient
MSF	Multi-Scale Fusion
pHash	Perceptual Hash
SE	Squeeze-and-Excitation
TTA	Test-Time Augmentation
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence

References

1. Food and Agriculture Organization of the United Nations. The Hidden Health Crisis: How Plant Diseases Threaten Global Food Security. FAO One Health Highlights, 2025. Available online: <https://www.fao.org/one-health/highlights/how-plant-diseases-threaten-global-food-security> (accessed on 20 July 2026).
2. Food and Agriculture Organization of the United Nations. *FAO's Plant Production and Protection Division; Divisional Overview Booklet; Technical Report*; FAO: Rome, Italy, 2022. [CrossRef]
3. Pathogens, precipitation and produce prices [Editorial]. *Nat. Clim. Change* **2021**, *11*, 635. [CrossRef]
4. Sun, J. Cross-dataset evaluation of deep learning models for plant pest and disease diagnosis. *bioRxiv* **2024**. [CrossRef]
5. Krishna, M.S.; Machado, P.; Otuka, R.I.; Yahaya, S.W.; Neves dos Santos, F.; Ihianle, I.K. Plant Leaf Disease Detection Using Deep Learning: A Multi-Dataset Approach. *J* **2025**, *8*, 4. [CrossRef]

6. Mustofa, S.; Munna, M.M.H.; Emon, Y.R.; Rabbany, G.; Ahad, M.T. A comprehensive review on Plant Leaf Disease detection using Deep learning. *arXiv* **2023**, arXiv:2308.14087.
7. Das, A.; Pathan, F.; Jim, J.R.; Kabir, M.M.; Mridha, M.F. Deep learning-based classification, detection, and segmentation of tomato leaf diseases: A state-of-the-art review. *Artif. Intell. Agric.* **2025**, *15*, 192–220. [\[CrossRef\]](#)
8. Yang, S.; Feng, Q.; Zhang, J.; Yang, W.; Zhou, W.; Yan, W. From laboratory to field: Cross-domain few-shot learning for crop disease identification in the field. *Front. Plant Sci.* **2024**, *15*, 1434222. [\[CrossRef\]](#)
9. Ahmmed, J.; Ahmed, F.; Shohan, R.H.; Rana, M.M.; Hasan, M. Fine-Tuned CNN-Based Approach for Multi-Class Mango Leaf Disease Detection. *arXiv* **2025**, arXiv:2510.05326.
10. Gopalan, K.; Srinivasan, S.; Pragma; Singh, M.; Mathivanan, S.K.; Moorthy, U. Corn leaf disease diagnosis: Enhancing accuracy with resnet152 and grad-cam for explainable AI. *BMC Plant Biol.* **2025**, *25*, 440. [\[CrossRef\]](#)
11. Alhammad, S.M.; Khafaga, D.S.; El-hady, W.M.; Samy, F.M.; Hosny, K.M. Deep learning and explainable AI for classification of potato leaf diseases. *Front. Artif. Intell.* **2025**, *7*, 1449329. [\[CrossRef\]](#)
12. Jayaprakash, V.; Akilesh, K.; Kumar, A.; Balamurugan, M.S.; Rajagopal, M.K. Cashew dataset generation using augmentation and RaLSGAN and a transfer learning based tinyML approach towards disease detection. *arXiv* **2023**, arXiv:2304.08766.
13. Tuteja, G.; Al-Yarimi, F.A.M.; Ikram, A.; Gupta, R.; Rehman, A.U.; Singh, J.; Delgado Noya, I.; Dzul Lopez, L.A. CNNAtLSTM: An attention-enhanced CNN–LSTM architecture for high-precision jackfruit leaf disease classification. *Front. Plant Sci.* **2026**, *16*, 1720471. [\[CrossRef\]](#)
14. Hossain, M.A.; Sakib, S.; Abdullah, H.M.; Arman, S.E. Deep learning for mango leaf disease identification: A vision transformer perspective. *Heliyon* **2024**, *10*, e36361. [\[CrossRef\]](#)
15. Karimanzira, D. Context-Aware Tomato Leaf Disease Detection Using Deep Learning in an Operational Framework. *Electronics* **2025**, *14*, 661. [\[CrossRef\]](#)
16. Rajendran, M.; Muruganantham, M. Tomato leaf disease classification using a vision transformer with cascaded group attention (ViT-CGA) for real-time decision support in precision agriculture. *J. Plant Dis. Prot.* **2026**, *133*, 91. [\[CrossRef\]](#)
17. Ergün, E. Attention-enhanced hybrid deep learning model for robust mango leaf disease classification via ConvNeXt and vision transformer fusion. *Front. Plant Sci.* **2025**, *16*, 1638520. [\[CrossRef\]](#)
18. Jia, S.; Wang, G.; Li, H.; Liu, Y.; Shi, L.; Yang, S. ConvTransNet-S: A CNN-Transformer Hybrid Disease Recognition Model for Complex Field Environments. *Plants* **2025**, *14*, 2252. [\[CrossRef\]](#)
19. Ahmmed, J.; Kabir, M.A.; Rehman, A.u.; Bermak, A. A lightweight hybrid CNN and transformer model for medicinal leaf disease classification with explainable AI. *Sci. Rep.* **2026**, *16*, 8243. [\[CrossRef\]](#)
20. Murugesan, S.; Chinnadurai, J.; Srinivasan, S.; Mathivanan, S.K.; Chandan, R.R.; Moorthy, U. Robust multiclass classification of crop leaf diseases using hybrid deep learning and Grad-CAM interpretability. *Sci. Rep.* **2025**, *15*, 29955. [\[CrossRef\]](#)
21. Das, S.; Prethija, G.; Sudha, C. LeafFusionNet: A hybrid deep learning approach for robust plant disease detection. *Front. Artif. Intell.* **2026**, *9*, 1773329. [\[CrossRef\]](#)
22. Pal, C.; Karmakar, S.; Mukherjee, I.; Chakrabarti, P.P. A lightweight and explainable CNN model for empowering plant disease diagnosis. *Sci. Rep.* **2025**, *15*, 30720. [\[CrossRef\]](#)
23. Hoang, P.H.; Phan, T.T.H. Cross-Architectural Mixture-of-Experts with Adaptive Soft Routing for Plant Leaf Disease Classification. *arXiv* **2026**, arXiv:2606.23441.
24. Yu, W.; Si, C.; Zhou, P.; Luo, M.; Zhou, Y.; Feng, J.; Yan, S.; Wang, X. MetaFormer Baselines for Vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 896–912. [\[CrossRef\]](#)
25. Karthikeyan, S.; Charan, R.; Narayanan, S.; Jani Anbarasi, L. Enhanced plant disease classification with attention-based convolutional neural network using squeeze and excitation mechanism. *Front. Artif. Intell.* **2025**, *8*, 1640549. [\[CrossRef\]](#)
26. Singh, B.; Sharma, R.P.; Dey, S. Interpretable Plant Leaf Disease Detection Using Attention-Enhanced CNN. *arXiv* **2025**, arXiv:2512.17864.
27. Yao, Y.; Deng, X.; Zhang, X.; Li, J.; Sun, W.; Zhang, G. Automatic visual recognition for leaf disease based on enhanced attention mechanism. *PeerJ Comput. Sci.* **2024**, *10*, e2365. [\[CrossRef\]](#)
28. Ilse, M.; Tomczak, J.M.; Welling, M. Attention-based Deep Multiple Instance Learning. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 2127–2136.
29. K, M.; Sankaranarayanan, S. Optimised hybrid late fusion deep learning model for cashew disease classification. *Front. Plant Sci.* **2026**, *17*, 1815033. [\[CrossRef\]](#)
30. Shah, S.S.A.; Saeed, F.; Raza, M.U.; Rehman, A.; Shaheryar, M.; Kim, I.M.; Yun, S.; Kang, J.M. PlantCLR: Contrastive self-supervised pretraining for generalizable plant disease detection. *Sci. Rep.* **2026**, *16*, 10550. [\[CrossRef\]](#)
31. Hoang, T.M.; Bui, V.H.; Nguyen, V.S.; Doan, D.T.; Dang, H.A.; Pham, A.T. A comprehensive evaluation of lightweight deep learning models for tomato disease classification on edge computing environments. *Sci. Rep.* **2026**, *16*, 12320. [\[CrossRef\]](#)
32. Duhan, S.; Gulia, P.; Gill, N.S.; Oliveira, T.A.; Singh, P.K. Evaluation of lightweight and efficient deep learning models for plant disease classification. *Discov. Internet Things* **2026**, *6*, 43. [\[CrossRef\]](#)

33. Arman, S.E.; Abdullah, H.M.; Sakib, S.N.; Saiem, R.M.; Asha, S.N.; Hasan, M.M.; Amin, S.B.; Abrar, S.M.M. SugarcaneShuffleNet: A Very Fast, Lightweight Convolutional Neural Network for Diagnosis of 15 Sugarcane Leaf Diseases. *arXiv* **2025**, arXiv:2508.17107.
34. Sun, J.; Feng, Z.; Han, J.; Xu, F.; Zhang, H.; Guo, Y. YOLO-PLNet: A lightweight real-time detection model for peanut leaf diseases based on edge deployment. *Front. Plant Sci.* **2025**, *16*, 1707501. [[CrossRef](#)]
35. Al Kafi, M.A.; Banshal, S.K.; Moni, R.; Aziz, A.L.; Aljuaid, M.; Bansal, R. LeafAI: Interpretable plant disease detection for edge computing. *PLoS ONE* **2026**, *21*, e0335956. [[CrossRef](#)]
36. Sun, H.; Fu, R.; Wang, X.; Wu, Y.; Al-Absi, M.A.; Cheng, Z.; Chen, Q.; Sun, Y. Efficient deep learning-based tomato leaf disease detection through global and local feature fusion. *BMC Plant Biol.* **2025**, *25*, 311. [[CrossRef](#)]
37. Srabon, T.H.; Siddiquee, S.M.T. *Multi-Crop Leaf Disease Dataset for Deep Learning-Based Classification and Explainable AI*; Mendeley Data, V1; Daffodil International University: Dhaka, Bangladesh, 2026. [[CrossRef](#)]
38. Shakib, M.M.H.; Mustofa, S.; Ahad, M.T. *MLD24: An Image Dataset for Mango Leaf Disease Detection*; Mendeley Data, V1; Daffodil International University: Dhaka, Bangladesh, 2024. [[CrossRef](#)]
39. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
40. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255. [[CrossRef](#)]
41. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [[CrossRef](#)]
42. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [[CrossRef](#)]
43. Zhou, K.; Yang, Y.; Qiao, Y.; Xiang, T. Domain Generalization with MixStyle. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
44. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
45. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
46. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 328–339. [[CrossRef](#)]
47. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13001–13008. [[CrossRef](#)]
48. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [[CrossRef](#)]
49. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. *mixup*: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
50. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6022–6031. [[CrossRef](#)]
51. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1195–1204.
52. Izmailov, P.; Podoprikin, D.; Garipov, T.; Vetrov, D.; Wilson, A.G. Averaging Weights Leads to Wider Optima and Better Generalization. In Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI), Monterey, CA, USA, 6–10 August 2018; pp. 876–885.
53. Wightman, R. PyTorch Image Models (timm), Version 1.0.28. Available online: <https://github.com/huggingface/pytorch-image-models> (accessed on 20 July 2026).
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
55. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [[CrossRef](#)]
56. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
57. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [[CrossRef](#)]

58. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995. [\[CrossRef\]](#)
59. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [\[CrossRef\]](#)
60. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [\[CrossRef\]](#)
61. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324. [\[CrossRef\]](#)
62. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; PMLR 97, pp. 6105–6114.
63. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q.V. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2815–2823. [\[CrossRef\]](#)
64. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing Network Design Spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10425–10433. [\[CrossRef\]](#)
65. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1577–1586. [\[CrossRef\]](#)
66. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11966–11976. [\[CrossRef\]](#)
67. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
68. Touvron, H.; Cord, M.; Jégou, H. DeiT III: Revenge of the ViT. In Computer Vision—ECCV 2022, Proceedings of the 17th European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2022; Volume 13684, pp. 516–533. [\[CrossRef\]](#)
69. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [\[CrossRef\]](#)
70. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, NSW, Australia, 6–11 August 2017; pp. 1321–1330.
71. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
72. McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [\[CrossRef\]](#)
73. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [\[CrossRef\]](#)
74. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [\[CrossRef\]](#)
75. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605. [\[CrossRef\]](#)
76. Windbiel-Rojas, K.; Messenger-Sikes, B. *Pest Notes: Sooty Mold*; UC ANR Publication 74108; University of California Agriculture and Natural Resources: Oakland, CA, USA, 2020. Available online: <https://ipm.ucanr.edu/home-and-landscape/sooty-mold/> (accessed on 1 August 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.