

A Localization-Aware Heterogeneous CNN Ensemble with Neural Meta- Fusion for Plant Disease Classification, with a Component Analysis on Laboratory and Field Images

Ehsan Gaballah

124220194696900@pg.cu.edu.eg

Cairo University

Research Article

Keywords: plant disease classification, leaf localization, Mask R-CNN, transfer learning, ensemble learning, stacked generalization, ablation study, within-dataset field-image evaluation

Posted Date: August 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10437236/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Deep convolutional networks now classify leaf images on curated benchmarks such as PlantVillage with accuracies close to the measurement ceiling of those datasets, which has shifted the open questions away from raw accuracy toward two under-reported issues: which components of a composite pipeline actually cause the result, and whether the components that matter on laboratory images are the same ones that matter on field photographs. We address both with a classification framework evaluated under a protocol that fixes every development decision before the independent test data are read. A Mask R-CNN stage localizes the dominant leaf, the accepted box is expanded by a validation-selected 5% margin and resized to a shared 224 by 224 input, and the same localized image is passed to ResNet50, InceptionV3, and MobileNetV2 together with an extractor that produces eighteen colour and shape descriptors. The forty-five class probabilities and eighteen descriptors form a sixty-three-dimensional input to a neural meta-classifier developed by three repetitions of stratified five-fold cross-validation. On a locked 3,101-image PlantVillage test partition of fifteen classes the framework reached 99.77% accuracy and 99.75% macro F1 with seven misclassifications, and a one-component-at-a-time ablation confirmed that every stage contributed. The same design was then trained and evaluated entirely within a separate thirteen-class PlantDoc field-image dataset, where it reached 90.03% accuracy and 89.40% macro F1. This is a within-PlantDoc experiment and not a controlled-to-field transfer test, so the figure measures how the pipeline behaves on field imagery rather than how a PlantVillage-trained model survives a domain shift. The central finding comes from running the identical ablation on both datasets: on clean images the ensemble breadth and descriptors provide the incremental gains, but under field conditions the ordering changes, and leaf localization and learned fusion become the decisive components. Removing localization cost 2.83 accuracy points and replacing the neural fusion with soft voting cost a further 2.08 points, the two largest effects on PlantDoc. The contribution is a controlled and transparent account of where each component of a localization-aware plant disease classifier earns its place, and of how that ordering shifts between laboratory and field acquisition.

1. Introduction

Plant diseases lower both the quantity and the quality of agricultural output, and their early recognition remains one of the practical bottlenecks in crop management. Field diagnosis still depends heavily on visual inspection by trained personnel, which is fast but subjective, or on laboratory confirmation, which is reliable but slow and difficult to scale. Automated image-based diagnosis has been studied for more than a decade as a way to narrow the gap between these two extremes, and the release of large annotated collections such as PlantVillage turned the problem into a standard supervised recognition task that deep convolutional networks handle well.

The consequence of that success is a saturation effect. On the PlantVillage-derived subsets that dominate the literature, reported accuracies now cluster in a narrow band close to 99%, and differences between competing methods often fall within the range of run-to-run variation. When several architectures all report values near the ceiling, a further tenth of a percentage point on the same clean

benchmark is weak evidence of a better method. Two limitations of this line of work matter more than the headline number. The first is that many studies report a single accuracy without a controlled account of which components of a composite pipeline actually caused the result. The second is that models trained and tested on laboratory images with uniform backgrounds do not necessarily hold up on photographs taken in the field, where background clutter, variable illumination, occlusion, and pose are the rule rather than the exception.

Transfer learning from ImageNet-pretrained convolutional networks became the default approach to plant disease classification after early demonstrations that fine-tuned deep models outperform handcrafted pipelines on PlantVillage [1, 2]. Comparative studies established that no single backbone is uniformly best across crops and conditions, and that residual, inception, and lightweight mobile families carry different inductive biases [3], which motivates ensembles that read several backbones rather than committing to one. Recent ensemble studies confirm that combining complementary networks improves accuracy over the best single model [4, 5], although many fuse features or average logits rather than training an explicit meta-learner. Stacked generalization offers the alternative of training a second-level model on the outputs of first-level learners [6, 7], and a neural meta-classifier can accept heterogeneous inputs, which allows deep probabilities and handcrafted descriptors to be fused in one model.

Colour, shape, and texture descriptors were the standard representation before deep learning and retain value as compact, interpretable signals. Vegetation indices such as Excess Green [8, 9] separate leaf tissue from background, and recent work shows that pairing handcrafted measurements with deep features can improve tomato leaf disease classification [10]. Restricting the classifier to the leaf region reduces the influence of background. Mask R-CNN [11] supplies an instance-level region that can be cropped and passed to a downstream classifier, as applied to apple leaf disease by Rehman et al. [12], and lesion-level analysis has been used to the same end [13]. On clean laboratory images the benefit of localization is modest because the background is already uniform, but its value is expected to grow on field images where the background carries most of the nuisance variation.

A growing body of evidence shows that high accuracy on controlled datasets does not transfer cleanly to field imagery. Xiang et al. [14] report that a PlantVillage-trained ResNet50 loses 67.7 accuracy points when it is applied directly to PlantDoc, and that standard mitigation techniques recover only part of that loss. PlantDoc was assembled to expose exactly this gap, since its images come from the open web with natural backgrounds [15]. The present work does not attempt that direct transfer. Instead it trains and evaluates the same pipeline design entirely within PlantDoc, and runs the same component ablation on both datasets, so that the comparison is between which components matter on clean images and which matter on field images, rather than between two absolute accuracies obtained under a domain shift.

This paper addresses both limitations directly. We build a classification framework that combines leaf localization, three heterogeneous convolutional backbones, an interpretable handcrafted branch, probability calibration, and a neural meta-classifier, and we evaluate it under a protocol that separates

every development decision from the independent test data. The framework is assessed in two settings. The first is the standard fifteen-class PlantVillage subset, where the question is how far the complete design can be pushed and how much each stage contributes on clean images. The second is a thirteen-class experiment trained and evaluated entirely within the PlantDoc field-image dataset, where the same components are examined under natural acquisition conditions rather than transferred from PlantVillage. Running the same ablation logic on both datasets makes it possible to state not just the final accuracy but where each component earns its place. The present study concerns the classification formulation of the problem; an object-detection formulation of tomato leaf disease diagnosis was addressed separately by the authors [16].

The specific contributions are as follows.

A localization-aware classification pipeline that supplies one validation-selected leaf-focused input to all deep and handcrafted branches, so that background variation is reduced before any feature is computed. The 5% crop margin and the 224 by 224 shared input size were chosen on validation data alone.

A sixty-three-input neural meta-classifier that fuses forty-five deep probabilities with eighteen colour and shape descriptors, developed through repeated stratified out-of-fold cross-validation with fold-internal feature scaling and evaluated once on a locked 3,101-image test partition.

A controlled component study on both datasets. On PlantVillage the study removes calibration, the handcrafted branch, and each backbone in turn. On PlantDoc it also removes the localization stage and replaces the neural fusion with soft voting, which lets the roles of localization and learned fusion be measured under field conditions.

A transparent comparison with recent work on the same subset that separates studies sharing the present locked-test protocol from studies whose evaluation design differs, so that raw accuracy is not treated as a single universal ranking.

2. Materials and Methods

The framework is a four-stage pipeline, shown in Fig. 1. A Mask R-CNN stage localizes the dominant leaf, the accepted box is expanded by a validation-selected margin and resized to a shared input, and the same localized image is passed to three convolutional backbones and to a handcrafted descriptor extractor. The forty-five deep probabilities and eighteen descriptors are concatenated into a sixty-three-dimensional vector and passed to a neural meta-classifier that returns the final class probabilities. Each stage is described in turn below, after the datasets and the fixed data split.

2.1 Datasets and their roles

Three public datasets support the classification framework. A PlantVillage-derived subset of 20,638 images across fifteen pepper, potato, and tomato classes provides the training and evaluation material for the classifier. A separate PlantVillage bounding-box export of 1,500 image-label pairs trains the Mask R-CNN leaf localizer. A PlantDoc field-image set of 1,097 images across the thirteen classes shared with the classification task supports a separate within-PlantDoc field-image evaluation. Table 1 summarizes the datasets and their roles.

Table 1
Datasets used in the study and their roles.

Dataset	Source	Scale used	Annotation	Role
PlantVillage subset	Kaggle [17]	20,638 images; 15 classes	One class label per image	Train and evaluate the classifier
PlantVillage box export	Roboflow [18]	1,500 image-label pairs	Leaf bounding boxes	Train the Mask R-CNN localizer
PlantDoc shared subset	Kaggle [19]; Singh et al. [15]	1,097 images; 13 shared classes	One class label per image	Within-PlantDoc field-image evaluation

2.2 Fixed data split and balanced training set

The 20,638 classification images were divided once with a stratified 70:15:15 ratio into 14,440 training, 3,097 validation, and 3,101 test images. The validation partition was then split further into 1,548 validation-selection images and 1,549 validation-meta images. The first subset was reserved for crop-padding selection, image-size selection, and backbone checkpoint selection. The second supported repeated out-of-fold development of the meta-classifier. The 3,101-image test partition was not read until every choice had been fixed.

Because the classes are unbalanced, a balanced training manifest of 2,000 effective entries per class was constructed. Classes below the target kept their original images and received deterministic augmented replicas, and the single class above the target was reduced without replacement. The final manifest contained 14,195 original entries and 15,805 augmented replicas, giving 30,000 effective training entries drawn from 14,195 unique source images. Augmentation was applied to the training manifest only, so no augmented view of a validation or test image ever entered training.

2.3 Stage I: Mask R-CNN leaf localization

Mask R-CNN [11] localizes the dominant leaf in each image. At a fixed acceptance confidence of 0.50 the localizer returned a box for all 20,638 classification images, so the full-image fallback rule was never triggered. Accepted-box confidences concentrated near one, with a mean of 0.99992 and a minimum accepted score of 0.56120. The selected box covered on average 72.26% of the original image area, with the middle half of images between 63.63% and 83.29% coverage, which indicates that localization removed peripheral background while retaining most of the visible leaf (Fig. 2). The accepted box was

expanded by the validation-selected 5% margin and resized to the shared input size before any feature was computed. The instance mask itself is not used by any later stage.

2.4 Selection of crop margin and shared input size

The crop margin and the shared input size were both chosen on the validation-selection subset alone. Four margins were screened at a common 128 by 128 input. Accuracy was identical at 0% and 5% margin, but macro F1 rose from 91.748% to 92.586% with the 5% margin, and larger margins reduced both measures, so 5% was retained. With the margin fixed, three input sizes were compared. Ensemble macro F1 rose from 92.586% at 128 pixels to 94.220% at 160 pixels and 96.665% at 224 pixels, and accuracy followed the same pattern to 97.222% at 224 pixels, so all branches used 224 by 224 localized images. Because these decisions used only validation data, the locked test result remains uncontaminated by hyperparameter search.

2.5 Stage II: convolutional probability branches

Three ImageNet-pretrained backbones were fine-tuned independently on the 30,000-entry balanced manifest: ResNet50 [20], InceptionV3 [21], and MobileNetV2 [22]. The three were chosen because they occupy different points on the accuracy-cost curve and carry different architectural priors, a residual design, a multi-branch inception design, and an inverted-residual lightweight design, which is the property an ensemble exploits. Each backbone received a task head of global average pooling, dropout of 0.30, and a fifteen-unit softmax. Training used a three-epoch frozen warm-up at learning rate 1e-3 followed by full fine-tuning for up to twelve epochs at 1e-5, with AdamW [23] and weight decay 1e-4. The best checkpoint was selected by validation macro F1, and validation loss governed early stopping and learning-rate reduction. Each backbone outputs a fifteen-class probability vector, so the three together contribute forty-five probability features.

2.6 Stage III: handcrafted colour and shape descriptors

The handcrafted branch operated on the same localized 224 by 224 image (Fig. 3). Fourteen colour descriptors were formed from the means and standard deviations of the R, G, B, H, S, V channels and the Excess Green vegetation index. Otsu thresholding [24] was applied to the Excess Green response to form a foreground mask, and the largest external contour was retained, from which solidity, circularity, aspect ratio, and area ratio were computed as four shape descriptors. The mask is an intermediate measurement rather than a learned segmentation output. Table 2 lists the eighteen-dimensional composition.

Table 2
Composition of the eighteen-dimensional handcrafted descriptor vector.

Feature group	Descriptors	Count
RGB statistics	Mean and standard deviation of R, G, B	6
HSV statistics	Mean and standard deviation of H, S, V	6
Vegetation index	Mean and standard deviation of Excess Green	2
Shape	Solidity, circularity, aspect ratio, area ratio	4
Total		18

2.7 Stage IV: neural meta-classifier

The forty-five deep probabilities and eighteen handcrafted descriptors were concatenated into a sixty-three-dimensional vector and passed to a compact neural meta-classifier, shown in Fig. 4. The network used a first block of 64 units with ReLU, batch normalization, and dropout 0.20, a second block of 32 units with ReLU, batch normalization, and dropout 0.10, and a fifteen-unit softmax output. Optimization used AdamW at learning rate 1e-3 with weight decay 1e-4. Development followed three repetitions of stratified five-fold cross-validation, giving fifteen out-of-fold rounds, and fold-specific feature scaling was fitted on each training portion only. After the architecture and settings were fixed, the system was evaluated once on the locked 3,101-image test partition. This out-of-fold protocol prevents the meta-classifier from ever fitting to predictions associated with the test images. Table 3 summarizes the settings.

Table 3
Architecture and training settings of the neural meta-classifier.

Setting	Value
Input dimension	63 (45 deep probabilities + 18 descriptors)
Hidden block 1	Dense 64, ReLU, BatchNorm, Dropout 0.20
Hidden block 2	Dense 32, ReLU, BatchNorm, Dropout 0.10
Output layer	Dense 15, Softmax
Optimizer / learning rate / weight decay	AdamW / 1e-3 / 1e-4
Development protocol	3 repetitions of stratified 5-fold cross-validation
Final evaluation	One pass on the locked 3,101-image test set
Random seed	42

2.8 Ablation design and evaluation protocol

The contribution of each stage was measured with a one-component-at-a-time ablation. On PlantVillage the study removed probability calibration, the handcrafted branch, and each backbone in turn, retraining each reduced variant under the same protocol and evaluating it on the same locked test partition, so that any difference reflects only the removed component. On PlantDoc the same logic was extended to remove the Mask R-CNN localization stage and to replace the neural fusion with soft voting, and the three backbones were also evaluated individually. The reporting used aggregate accuracy and balanced accuracy together with macro-averaged precision, recall, and F1, because macro measures reveal whether the aggregate is carried by a single large class. The PlantDoc analysis was developed and evaluated within its own protocol rather than by transferring the PlantVillage-trained model, so it is a within-PlantDoc evaluation and not a direct cross-dataset transfer experiment.

3. Results

3.1 Individual backbone performance on PlantVillage

The three backbones converged quickly during fine-tuning and reached similar validation accuracies (Fig. 5). On the independent test partition ResNet50 was the strongest single model at 98.613% accuracy and 98.583% macro F1 with 43 errors. MobileNetV2 and InceptionV3 followed at 97.872% and 97.646% accuracy with 66 and 73 errors. MobileNetV2 reached this level with 2.28 million parameters, roughly one tenth of the count of the two heavier networks, which supports its inclusion as an efficient complementary branch even though it is not strong enough to carry the task alone. Table 4 reports the individual results.

Model	Parameters	Test accuracy	Balanced accuracy	Macro F1	Errors
InceptionV3	21,833,519	97.646%	97.232%	97.290%	73
MobileNetV2	2,277,199	97.872%	97.735%	97.772%	66
ResNet50	23,618,447	98.613%	98.471%	98.583%	43

Table 4. Independent-test results of the three localized-input backbones.

3.2 Full meta-fusion model on PlantVillage

The complete sixty-three-input meta-classifier reached 99.77% accuracy on the locked test partition, correctly classifying 3,094 of 3,101 images with seven errors. Balanced accuracy was 99.72%, macro precision 99.79%, macro recall 99.72%, macro F1 99.75%, and weighted F1 99.77%. The close agreement between overall accuracy and weighted F1 indicates that the result was not driven by a single large class, and the macro measures confirm consistent performance across the fifteen categories. The fusion improved on the best single backbone by more than one accuracy point, which reduced the error count from 43 to 7. Table 5 reports the aggregate measures.

Measure	Result
Test accuracy	99.77%
Correct predictions	3,094 of 3,101
Test errors	7
Balanced accuracy	99.72%
Macro precision	99.79%
Macro recall	99.72%
Macro F1	99.75%
Weighted F1	99.77%

Table 5. Aggregate performance of the full 63-input meta-fusion model on the independent test set.

3.3 Class-wise behaviour

The class-wise report showed near-ceiling performance across all fifteen categories. Several classes were classified without error, including pepper healthy, potato healthy, tomato mosaic virus, tomato yellow leaf curl virus, and tomato healthy. The smallest class, potato healthy with 23 test images, retained perfect classification despite its limited support. The seven residual errors were isolated rather than concentrated in a single category, and each class-level F1 remained close to the upper end of the scale, which is consistent with the macro F1 of 99.75%. The confusion analysis in Figures 6 and 7 confirmed that off-diagonal mass was sparse and confined to a few visually similar disease pairs, with no broad systematic failure.

3.4 PlantVillage ablation study

Every removed component reduced accuracy and macro F1, which supports the complete design over any single stage. Among the backbones, removing ResNet50 caused the largest decline, 0.38 accuracy points and 0.40 macro F1 points, consistent with its status as the strongest individual model. Removing InceptionV3 and MobileNetV2 caused smaller but consistent drops, which indicates that the three backbones carry complementary probability information rather than redundant signals. Removing the handcrafted branch cost 0.16 accuracy points, direct evidence of a modest but measurable contribution from the explicit colour and shape representation. Calibration produced the smallest single gain. These magnitudes are small in absolute terms: on a 3,101-image test set a difference of 0.10 accuracy points corresponds to roughly three images, and the full ablation spans about twelve images between the strongest and the weakest reduced variant. Because each configuration was trained under a single seed, the individual gaps on PlantVillage should be read as a consistent ranking of contributions rather than as

precisely estimated effect sizes; the ranking is stable and every component helps, but the exact point values sit within the range where a few images can move them. Table 6 reports the ablation.

Configuration	Accuracy (%)	Macro F1 (%)	Acc. drop (pp)	F1 drop (pp)
Full proposed model	99.77	99.75	—	—
Without calibration	99.66	99.63	-0.11	-0.12
Without handcrafted descriptors	99.61	99.59	-0.16	-0.16
Without MobileNetV2	99.55	99.51	-0.22	-0.24
Without InceptionV3	99.48	99.44	-0.29	-0.31
Without ResNet50	99.39	99.35	-0.38	-0.40

Table 6. One-component-at-a-time ablation on the locked PlantVillage test partition.

3.5 Comparison with recent studies on the same subset

To position the result, the closest-protocol comparison was restricted to peer-reviewed studies that used the same fifteen-class PlantVillage subset of 20,638 images. Under this restriction the framework reached 99.77% accuracy, above the 99.64% reported by Hasani et al. [25] and the 98.41% reported by Tiwari et al. [26]. These differences are small and should be read with care, because architecture, checkpoint selection, augmentation, repeated-run reporting, and test construction all differ across studies. The comparison is deliberately restricted to studies that share the same fifteen-class subset and a comparable held-out test protocol, so that the numbers are placed on a common footing rather than ranked across incompatible designs. Several other works on the same subset used materially different designs, including two 80/20 splits [5, 27], a validation accuracy reported in place of a test accuracy [28], federated client evaluation [29], and external validation after training on a different dataset [30], so their reported values are treated as contextual benchmarks rather than as a direct ranking. What distinguishes the present evaluation is the locked 3,101-image test partition, the complete class-wise reporting, and the two-dataset ablation, all of which make the reported figures traceable to a stated protocol. Table 7 lists the closest-protocol comparisons.

Study	Setting	Reported accuracy
This work	Localization-aware three-backbone fusion with 18 descriptors and neural meta-classifier; locked 3,101-image test set	99.77%
Hasani et al. [25]	Comparative deep-learning evaluation on the same fifteen-class subset	99.64%
Tiwari et al. [26]	Deep-learning crop disease classifier on the same subset	98.41%

Table 7. Closest-protocol comparisons with recent peer-reviewed studies on the same fifteen-class PlantVillage subset.

3.6 Within-PlantDoc field-image evaluation

The controlled benchmark measures how far the design can be pushed, but it cannot show whether the same design survives field acquisition. To test this, the framework was evaluated within a separate thirteen-class PlantDoc experiment on 1,097 field images covering two pepper classes, two potato classes, and nine tomato classes. Unlike the uniform PlantVillage images, these photographs carry cluttered natural backgrounds, variable illumination, occlusion, scale changes, and heterogeneous leaf presentation, and their class distribution is highly uneven, ranging from 2 images for tomato spider mites to 150 for tomato Septoria leaf spot. Fig. 8 shows one representative image from each of the thirteen shared classes and makes the domain difference visible. The PlantDoc images were never merged with PlantVillage training data, so the experiment is a within-PlantDoc evaluation rather than a direct PlantVillage-to-PlantDoc transfer. The extreme imbalance should be kept in mind when reading the macro measures on this dataset: the smallest classes are supported by only a handful of test images, so a single prediction can move a per-class F1 by a large amount, and the PlantDoc macro figures carry correspondingly wide uncertainty.

The complete model reached 90.03% accuracy and 89.40% macro F1 on this field material. The absolute accuracy is about ten points below the PlantVillage level, but the two numbers are not directly comparable, because the PlantDoc model was trained on PlantDoc rather than transferred from PlantVillage. The gap therefore reflects the intrinsic difficulty of field images under a matched training regime, and it is much smaller than the collapse reported for direct controlled-to-field transfer, where Xiang et al. [14] measured a 67.7-point loss for a PlantVillage-trained ResNet50 applied to PlantDoc. The within-PlantDoc figure and the transfer figure answer different questions and should not be read on the same scale. The neural fusion again improved on every individual backbone, exceeding ResNet50, InceptionV3, and MobileNetV2 by 2.19, 3.08, and 3.72 accuracy points, a wider margin than on the clean benchmark, which is the first sign that fusion contributes more when the inputs are noisier. Table 8 reports the individual and fused accuracies.

Model	Accuracy (%)
ResNet50	87.84
InceptionV3	86.95
MobileNetV2	86.31
Neural fusion (full model)	90.03

Table 8. Individual backbone and fused accuracy on the PlantDoc field-image experiment.

The most informative result is the ablation, because it can be compared directly against the PlantVillage ablation of Section 3.4. On the clean benchmark the largest marginal contributors were the deep backbones, with ResNet50 first and the handcrafted branch and calibration adding smaller gains. On PlantDoc the ordering changes. Removing Mask R-CNN localization now produces the largest single effect, a 2.83-point accuracy drop and a 3.10-point macro F1 drop, because localization removes the cluttered background that dominates nuisance variation in field images. Replacing the neural fusion with soft voting causes the second-largest decrease of 2.08 accuracy points, which shows that a learned combination of the branches is worth more than a fixed average once the individual predictions disagree more often. The deep backbones still matter, with ResNet50 the largest among them, but they are no longer the top of the list. Table 9 and Fig. 9 report the full ablation.

Configuration	Accuracy (%)	Macro F1 (%)	Acc. change (pp)	F1 change (pp)
Full model	90.03	89.40	—	—
Without Mask R-CNN	87.20	86.30	-2.83	-3.10
Soft voting instead of ANN	87.95	87.05	-2.08	-2.35
Without ResNet50	88.45	87.60	-1.58	-1.80
Without handcrafted features	88.76	87.95	-1.27	-1.45
Without InceptionV3	88.98	88.15	-1.05	-1.25
Without MobileNetV2	89.22	88.40	-0.81	-1.00
Without calibration	89.55	88.85	-0.48	-0.55

Table 9. One-component-at-a-time ablation on the PlantDoc field-image experiment.

4. Discussion

The two evaluations answer different questions. On PlantVillage the framework reaches 99.77% accuracy with seven errors, which places it among the strongest reported results on that subset while remaining fully traceable to a locked protocol. Because the whole field operates near the ceiling of this benchmark, the more informative outcome is the ablation, which shows a consistent ordering in which ResNet50 contributes most among the backbones and the handcrafted branch and calibration add smaller but real gains. The value of the clean-benchmark result is therefore as much in the controlled attribution as in the accuracy figure.

The PlantDoc result is where the design earns its structure. Accuracy falls to 90.03%, and the ablation reorders the components so that leaf localization and learned fusion become the two largest contributors. This is the expected consequence of moving from uniform laboratory backgrounds to field

photographs. When the background is clean, localization removes little useful nuisance and the ensemble does most of the work. When the background is cluttered, localization removes the dominant source of variation before any feature is computed, and a learned meta-classifier is better than a fixed average at discounting a branch that has been misled by the scene. The same reordering is what the cross-domain literature would predict, since the field setting is where background and acquisition, rather than the choice of backbone, dominate the error [14].

Several limitations bound these claims. The test partitions are small in absolute terms, 3,101 images for the clean benchmark and 1,097 for the field experiment, so on PlantVillage a difference below roughly one third of a percentage point corresponds to only a handful of images and is reported as part of a consistent ranking rather than as a precisely estimated effect. Each configuration was trained under a single random seed, so the exact ablation magnitudes are not accompanied by run-to-run variance; the ordering of the components is stable across the two datasets, and it is the ordering, not the third-decimal value, that carries the argument. The PlantDoc analysis was trained and evaluated entirely within PlantDoc, so it characterizes how the pipeline behaves on field images but does not establish direct transfer of the PlantVillage-trained model to a new domain; the collapse measured for that direct transfer by Xiang et al. [14] is far larger than any gap reported here and marks the separate problem that this study does not attempt to solve. Efficiency is discussed only through parameter counts, and no latency, memory, or energy measurement on target hardware was carried out, so no real-time claim is made. Finally, the datasets are public and the class definitions were taken from their providers, which fixes the label space and its imperfections.

5. Conclusion

This paper presented a localization-aware classification framework that combines Mask R-CNN leaf localization, three heterogeneous convolutional backbones, an eighteen-descriptor handcrafted branch, probability calibration, and a sixty-three-input neural meta-classifier, and evaluated it under a protocol that fixes every decision before the independent test data are read. On the fifteen-class PlantVillage subset the framework reached 99.77% accuracy and 99.75% macro F1 with seven errors, and a controlled ablation confirmed that every stage contributed, with ResNet50 the largest among the backbones. On a separate thirteen-class PlantDoc field-image experiment the same design reached 90.03% accuracy, and its ablation showed that leaf localization and learned fusion, rather than the ensemble breadth or the descriptors, are the components that hold performance together under natural acquisition conditions. The practical message is that the parts of a composite pipeline that matter most depend on the operating conditions, and that a field-image ablation is needed to see this. Future work will pursue a direct cross-domain transfer evaluation from controlled to field data, a deployment benchmark that measures latency and memory on target hardware, and an extension of the framework to additional crops and to a larger and more diverse field dataset.

Declarations

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

Ethics approval:

Not applicable.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Ehsan Zeinhom Gaballah: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft, Writing - review and editing. Zeinab Ezz Alarab: Supervision, Methodology, Validation, Writing - review and editing. Mervat Hassan Geith: Supervision, Methodology, Validation, Writing - review and editing.

Data Availability

The image data supporting the findings of this study are derived from the publicly available PlantVillage dataset. The classification subset used here (20,638 images spanning fifteen pepper-bell, potato, and tomato classes) was obtained from PlantVillage, which is openly accessible at <https://www.kaggle.com/datasets/emmarex/plantdisease>.

References

1. Mohanty, S. P., Hughes, D. P., & Salathe, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419. <https://doi.org/10.3389/fpls.2016.01419>

2. Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145, 311–318. <https://doi.org/10.1016/j.compag.2018.01.009>
3. Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, 161, 272–279. <https://doi.org/10.1016/j.compag.2018.03.032>
4. Sharma, J., Al-Huqail, A. A., Almogren, A., Doshi, H., Jayaprakash, B., Bharathi, B., et al. (2025). Deep learning based ensemble model for accurate tomato leaf disease classification by leveraging ResNet50 and MobileNetV2 architectures. *Scientific Reports*, 15, 13904. <https://doi.org/10.1038/s41598-025-98015-x>
5. Aderinwale, D. A., & Sah, M. (2025). Plant disease classification using an ensemble learning model of convolutional neural networks and vision transformers. Proceedings of the 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA), pp. 1–10. <https://doi.org/10.1109/ICHORA65333.2025.11017080>
6. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
7. Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1), 49–64. <https://doi.org/10.1007/BF00117832>
8. Woebbecke, D. M., Meyer, G. E., Von Bargen, K., & Mortensen, D. A. (1995). Color indices for weed identification under various soil, residue and lighting conditions. *Transactions of the ASAE*, 38(1), 259–269. <https://doi.org/10.13031/2013.27838>
9. Meyer, G. E., & Neto, J. C. (2008). Verification of color vegetation indices for automated crop imaging applications. *Computers and Electronics in Agriculture*, 63(2), 282–293. <https://doi.org/10.1016/j.compag.2008.03.009>
10. Bouni, M., Hssina, B., Douzi, K., & Douzi, S. (2024). Synergistic use of handcrafted and deep learning features for tomato leaf disease classification. *Scientific Reports*, 14, 26822. <https://doi.org/10.1038/s41598-024-71225-5>
11. He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969. <https://doi.org/10.1109/ICCV.2017.322>
12. Rehman, Z. U., Khan, M. A., Ahmed, F., Damasevicius, R., Naqvi, S. R., Nisar, W., & Javed, K. (2021). Recognizing apple leaf diseases using a novel parallel real-time processing framework based on Mask RCNN and transfer learning. *IET Image Processing*, 15(10), 2157–2168. <https://doi.org/10.1049/ipr2.12183>
13. Barbedo, J. G. A. (2019). Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering*, 180, 96–107. <https://doi.org/10.1016/j.biosystemseng.2019.02.002>
14. Xiang, K., Shi, D., & Zhu, X. (2026). Quantifying the reliability gap in cross-domain plant disease classification: Benchmarking the limited efficacy of standard mitigation techniques under

- controlled-to-field shift. *Frontiers in Plant Science*, 17, 1826962.
<https://doi.org/10.3389/fpls.2026.1826962>
15. Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. (2020). PlantDoc: A dataset for visual plant disease detection. *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pp. 249–253.
<https://doi.org/10.1145/3371158.3371196>
 16. Gaballah, E. Z., Ezz Alarab, Z., & Geith, M. H. (2025). Tomato leaf disease detection using YOLOv12. *Journal of Intelligent Computing and Data Science*, 1(1), 1–24.
<https://doi.org/10.21608/jicds.2025.434079.1001>
 17. Emmarex (2018). PlantVillage Dataset [Data set]. Kaggle.
<https://www.kaggle.com/datasets/emmarex/plantdisease>
 18. zkamlasi kamlasi (2023). plantvillage-dataset [Data set]. Roboflow Universe.
<https://universe.roboflow.com/zkamlasi-kamlasi-hj4wj/plantvillage-dataset>
 19. Uddin, A. H. (2021). PlantDoc Dataset [Data set]. Kaggle. Original images derived from the PlantDoc repository of Singh et al. [15].
 20. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
<https://doi.org/10.1109/CVPR.2016.90>
 21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
 22. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
 23. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations*.
 24. Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems Man and Cybernetics*, 9(1), 62–66. <https://doi.org/10.1109/TSMC.1979.4310076>
 25. Hasani, Z., Fondaj, J., Bytyqi, D., & Misini, S. (2026). Comparative evaluation of deep learning models for plant disease classification with edge-aware performance analysis. *PLOS ONE*, 21(6), e0349901.
<https://doi.org/10.1371/journal.pone.0349901>
 26. Tiwari, U., Rai, V., & Rai, A. K. (2026). Smart disease detector in crops: A deep learning-based approach for automated plant disease identification. *Proceedings of IMPACT-2026, DMPedia LNCSE*, pp. 581–591. <https://doi.org/10.65890/dmp-incse.IMPACT26.160>
 27. Srivastava, A. K., Tiwari, A. M., & Singh, A. (2023). Fast AI and deep learning based optimized plant leaf disease detection. *EasyChair Preprint No.* 9844.
 28. Tabbakh, A., & Barpanda, S. S. (2023). A deep features extraction model based on the transfer learning model and vision transformer TLMViT for plant disease classification. *Ieee Access* :

29. Janga, R., Dave, R., & Bhavsar, M. (2025). AgriFedNet: Privacy-preserving plant pathology with MobileNetV3-Small. *Journal of Computing Sciences in Colleges*, 41(4), 101–111.
30. Subhan, M., Rana, N., Qamar, F., Jeribi, F., Alhazmi, A. H., & Tahir, A. (2026). Crop disease detection using EfficientNetB0 deep learning approach for precision agriculture. *Discover Computing*, 29, 25. <https://doi.org/10.1007/s10791-025-09881-y>

Figures

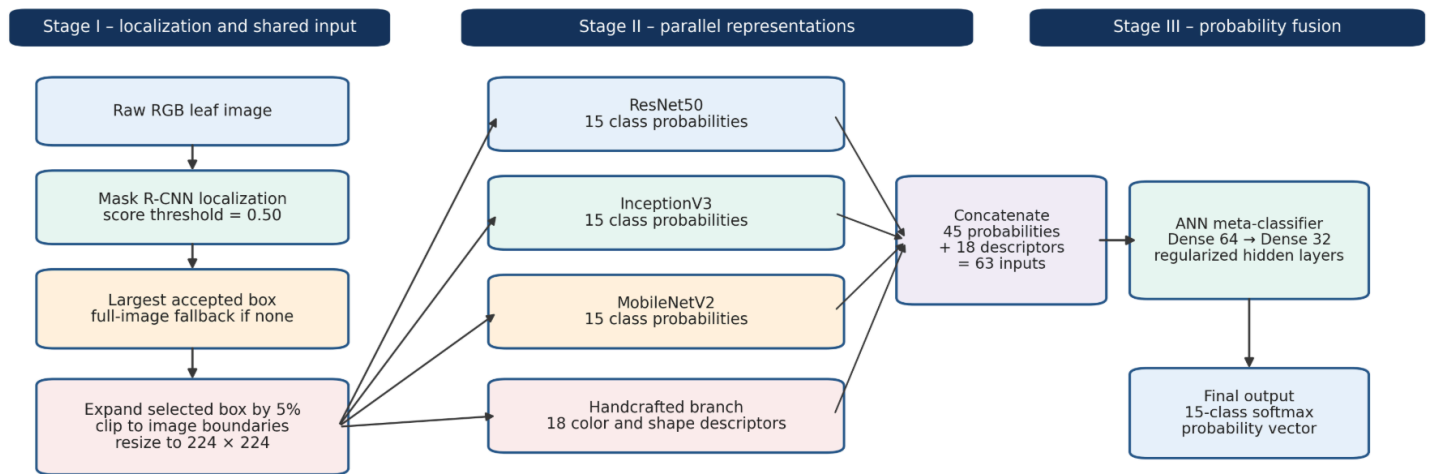


Figure 1

The four-stage classification pipeline. Mask R-CNN localization is followed by three convolutional branches and a handcrafted branch that share one 224 by 224 localized input, and their outputs are fused by a neural meta-classifier.

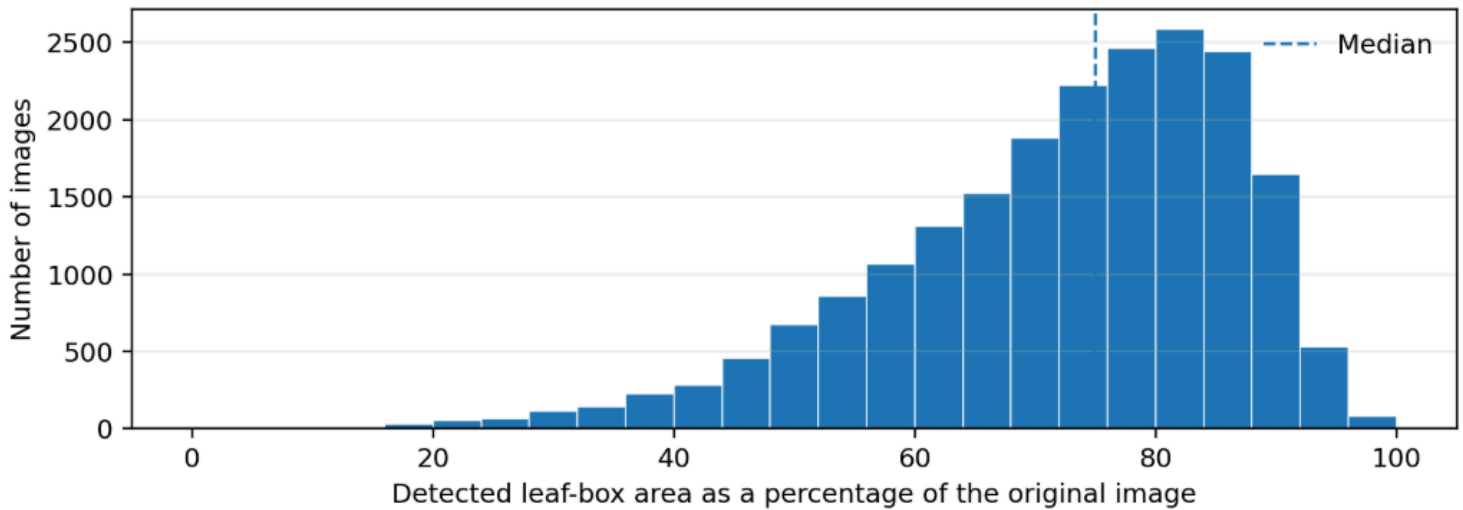
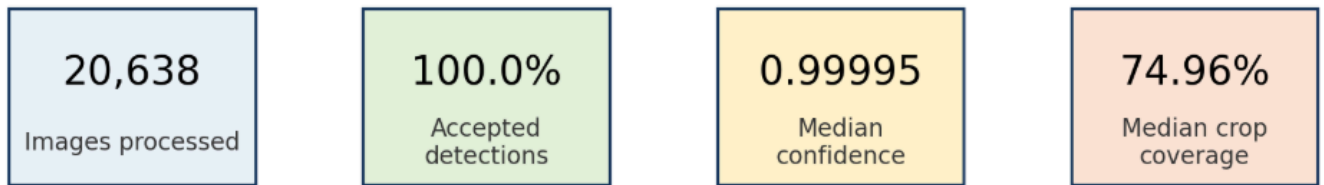


Figure 2

Localization behaviour across the full classification dataset: accepted-detection rate, confidence summary, and detected leaf-box area coverage.

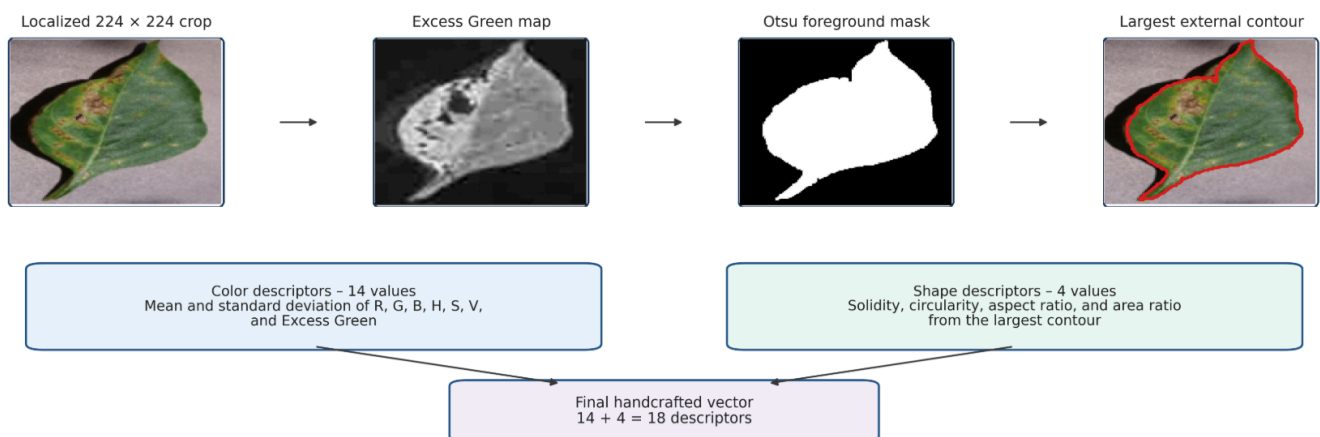
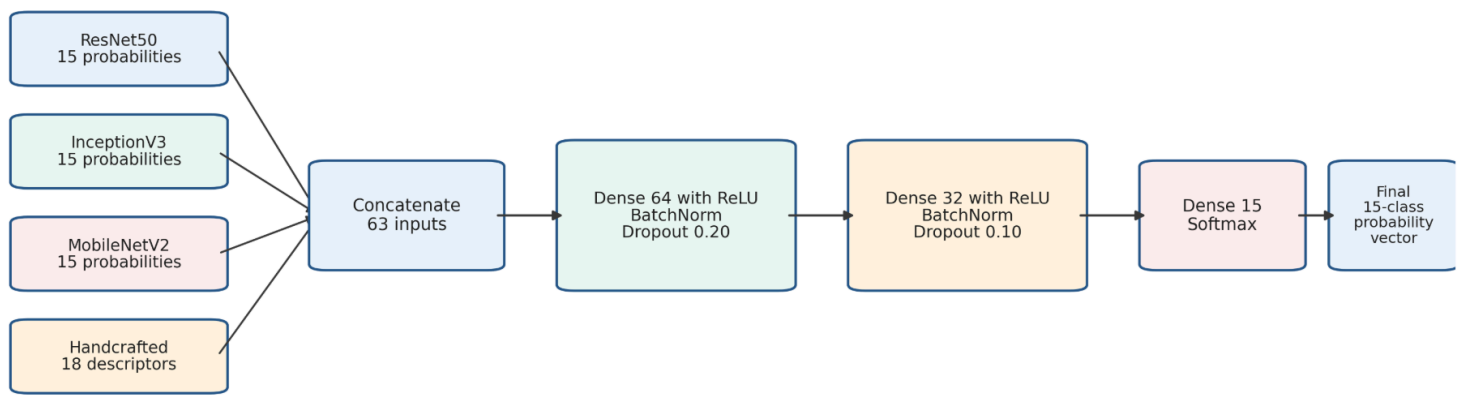


Figure 3

Extraction of the eighteen handcrafted descriptors, fourteen colour and four shape, from the shared localized 224 by 224 image.



Meta-development: 1,549 images • repeated 3 × 5 stratified folds • fold-specific scaling and balancing • 15 test probability matrices averaged

Figure 4

Neural meta-classifier and the repeated stratified five-fold out-of-fold development protocol. The sixty-three-dimensional input combines forty-five deep probabilities with eighteen handcrafted descriptors.

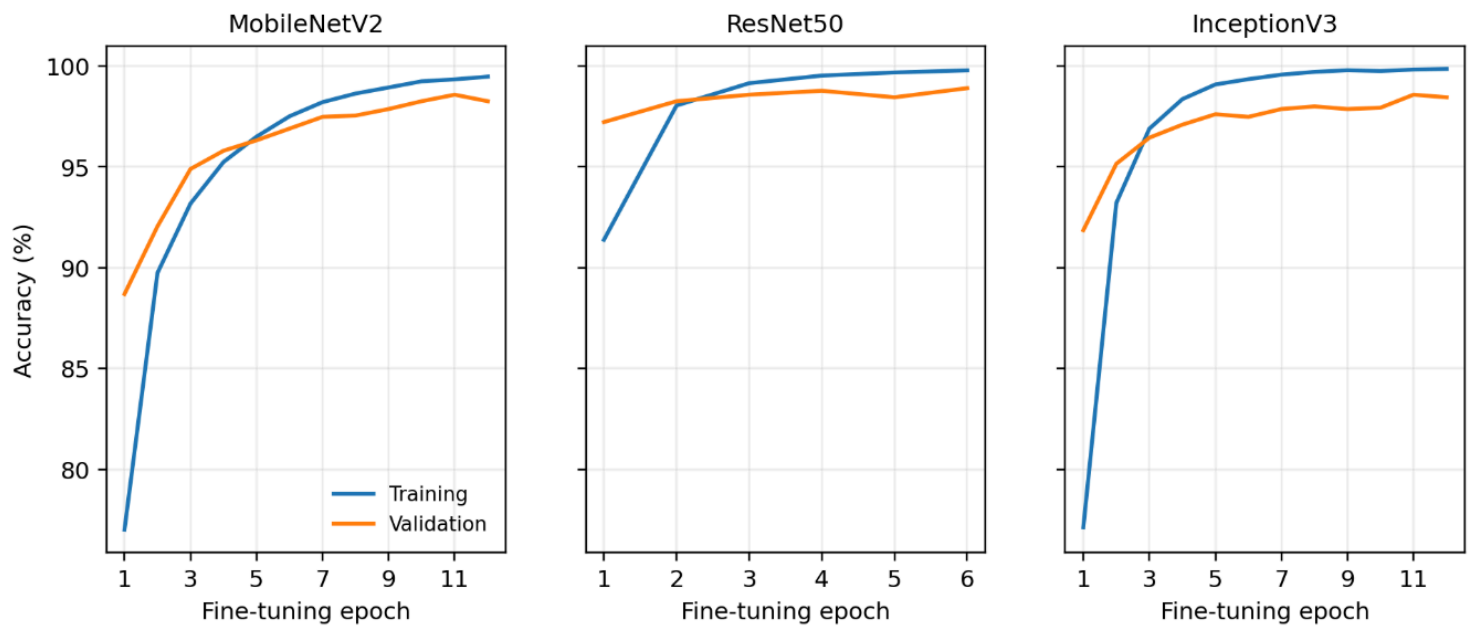


Figure 5

Training and validation accuracy during the fine-tuning stage of MobileNetV2, ResNet50, and InceptionV3 on localized 224 by 224 inputs.

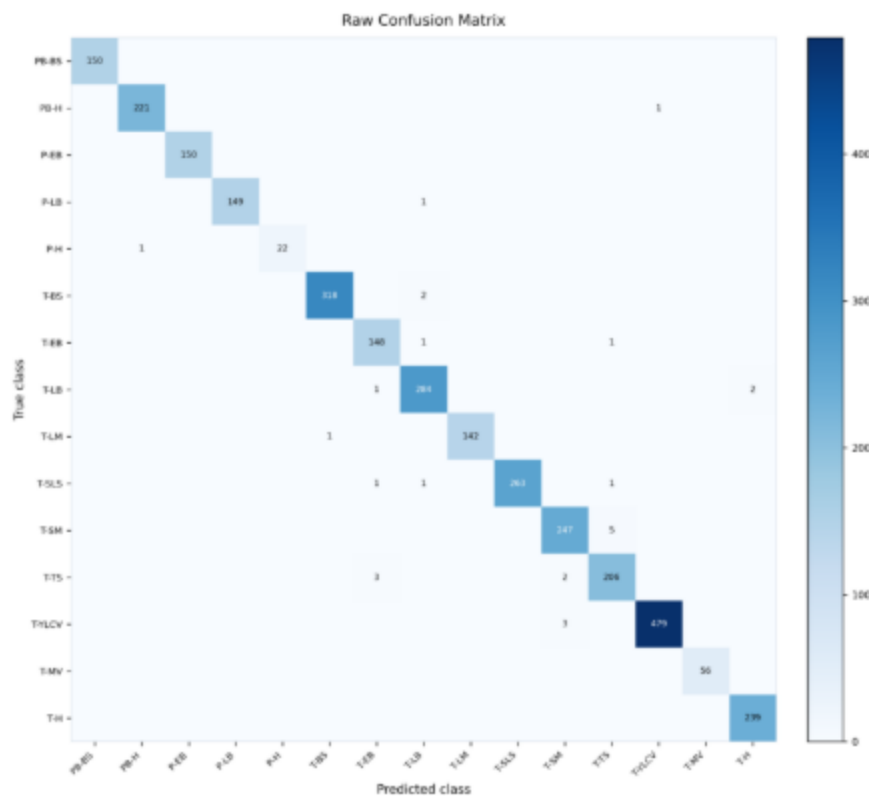


Figure 6

Raw confusion matrix of the full sixty-three-input meta-classifier on the 3,101-image independent test set. Rows are true classes and columns are predicted classes; only nonzero cells are labelled.

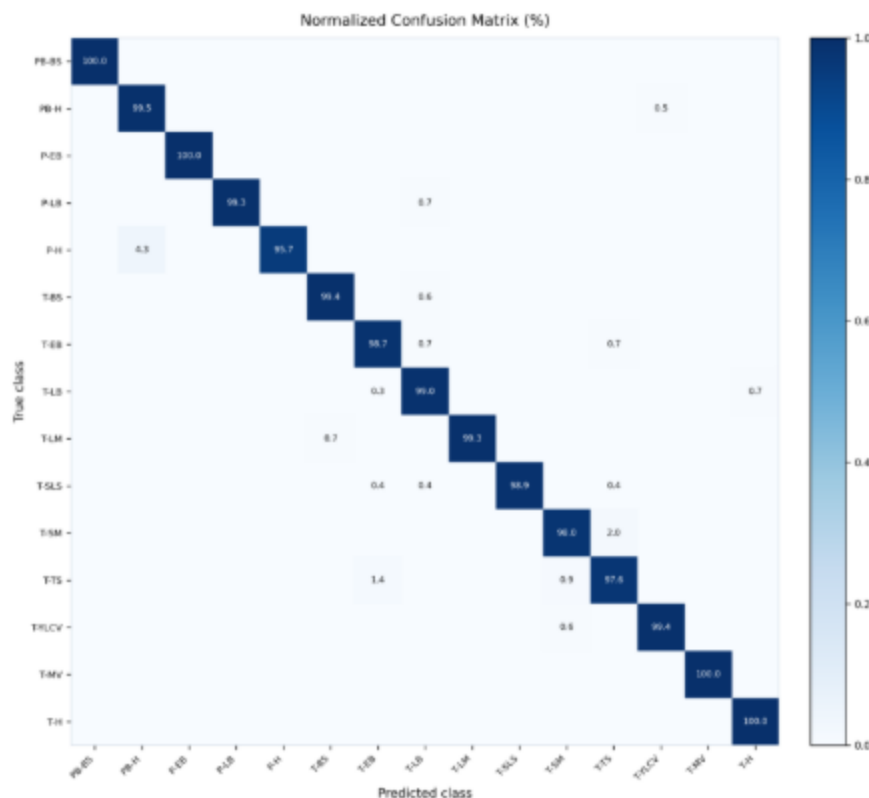


Figure 7

Row-normalized confusion matrix on the independent test set. Each row sums to 100% and each cell reports the percentage of a true class assigned to the corresponding predicted class.

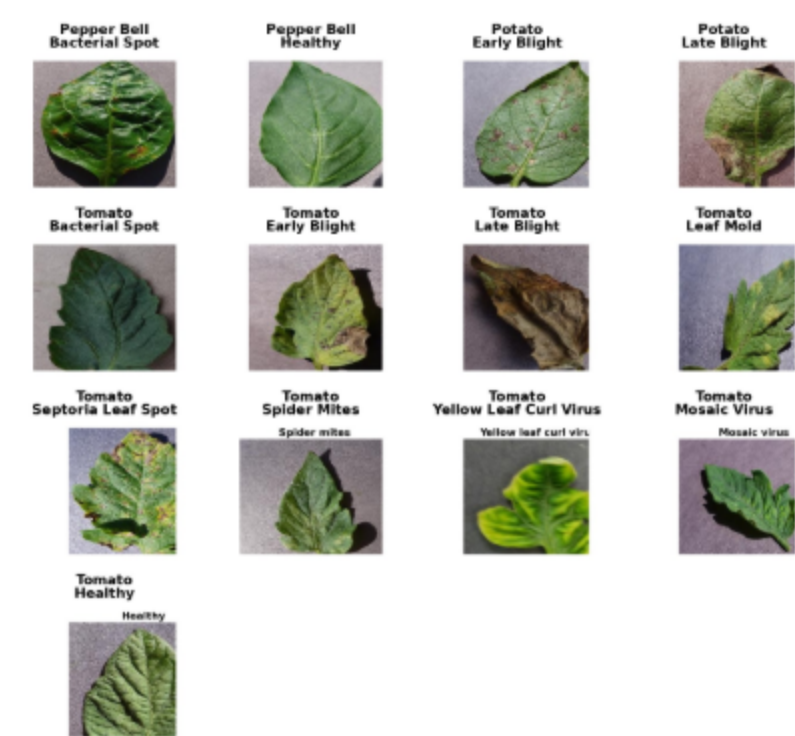


Figure 8

One representative field image from each of the thirteen shared PlantDoc classes. Natural backgrounds, variable illumination, and uncontrolled pose contrast with the uniform PlantVillage acquisition and motivate the localization stage.

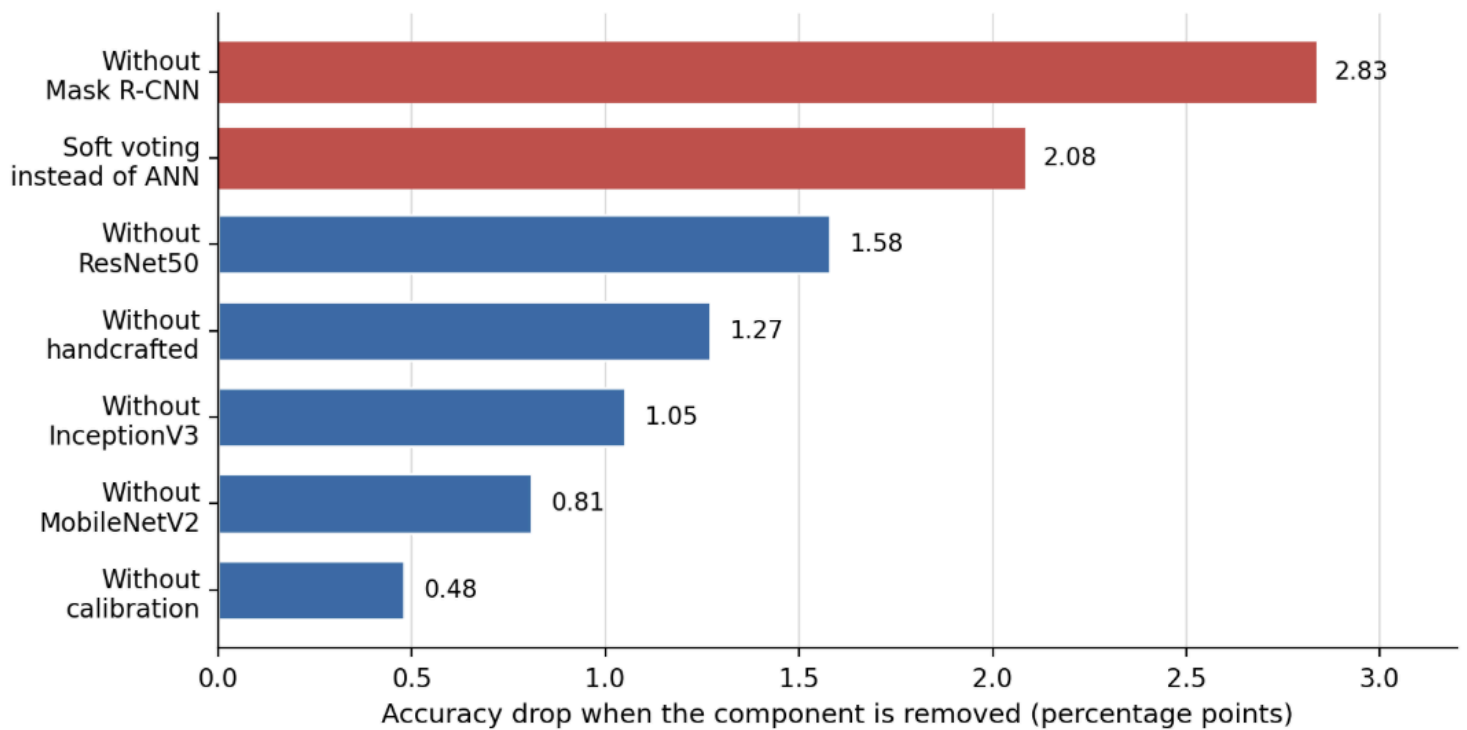


Figure 9

Accuracy drop when each component is removed from the PlantDoc field-image model. Leaf localization and learned fusion (highlighted) are the two largest contributors, reversing the ordering seen on the clean PlantVillage benchmark.