

Explainable Deep Learning for Multi-Crop Disease Classification with Grad-CAM-Based Visual Interpretation

Harish Kumar Mittal¹

Professor, Department of Computer Science
and Engineering, UIE,
Chandigarh University, Mohali, Punjab,
India
harish_mittal@yahoo.co.in

Pratibha Mittal²

Student, Department of CSE,
Indira Gandhi Delhi Technical University
for Women (IGDTUW), Delhi, India
pratibhartk1@gmail.com

Sunita³,

Professor and Head CSE,
St. Andrew's Institute of Technology and
Management Gurugram, India
sunitabhardwaj.cse@saitm.ac.in

Parvinder Bangar⁴

Professor – cum – Director, Greater Noida
Institute of Technology, Greater Noida,
Uttar Pradesh, India
parvinder.bangar@gmail.com

Madhav Mittal⁵

Student, Department of CSE,
Netaji Subhas University of Technology
Delhi, India
madhavrtk@gmail.com

Neeraj Sharma⁶

Department of CSE,
Lovely Professional University
Jalandhar, Punjab, India
neerajkirti@gmail.com

Abstract—Reliable plant-disease diagnosis on resource-constrained devices requires computational efficiency, calibrated confidence and interpretable predictions. This study fine-tunes three ImageNet-pretrained backbones—MobileNetV3-Small, MobileNetV2 and EfficientNet-B0—end to end on PlantVillage (54,305 images; 38 classes) under a fixed stratified split, each with three random seeds. All three exceed 99.7% test accuracy; EfficientNet-B0 is significantly the most accurate at 0.9984 ± 0.0002 , yet its margin over MobileNetV3-Small is 0.10 percentage points against 2.6 times the parameters and 6.8 times the floating-point cost, so accuracy no longer discriminates usefully and the deployment choice rests on cost. Temperature scaling leaves accuracy untouched while reducing expected calibration error from 0.0193 to 0.0009, and a calibrated three-seed ensemble lowers it further to 0.0005 while making 11 errors in 8,146 images. The models are near-insensitive to blur, sensor noise, JPEG compression and brightness shifts, the worst costing 0.14 percentage points. Grad-CAM attribution inside the leaf (0.513) is well above a random heatmap (0.379) and somewhat below a centre-weighted Gaussian (0.548), indicating that part of the localisation follows the centred framing of the dataset. Applied to PlantDoc field photographs without adaptation, accuracy is 15–20%, which marks domain adaptation as the main remaining step toward deployment. MobileNetV3-Small offers the best accuracy-to-cost balance for on-device use.

Index Terms—plant disease classification, explainable artificial intelligence, Grad-CAM, lightweight CNN, transfer learning, domain shift, precision agriculture

I. INTRODUCTION

Plant pests and diseases reduce global crop yields by an estimated 20 to 40% annually [1], threatening food security and agricultural livelihoods. Because many diseases produce visible leaf symptoms, image-based diagnosis can support earlier intervention. Although convolutional neural networks (CNNs) can achieve very high accuracy on controlled leaf-image datasets [2], such results do not establish robustness under field conditions.

Two limitations complicate practical deployment. First, computationally intensive models can be difficult to execute on resource-constrained mobile or edge hardware. Second, conventional CNN predictions provide limited insight into the image evidence used for classification. Without a visual rationale, agronomists and end users cannot readily determine whether a prediction is based on a lesion, the leaf surface, or irrelevant background. The same pairing of a compact model with an interpretable output has been pursued in other decision-support settings, where a lightweight explainable framework was shown to deliver actionable transparency without incurring heavy computational overhead [15].

This paper addresses efficiency, reliability and interpretability within a single empirical evaluation. Its contributions are: (i) a controlled comparison of three compact CNN backbones fine-tuned end to end under one protocol, with repeated seeds, confidence intervals and significance tests for every model rather than single-run point estimates; (ii) a calibration analysis showing that high accuracy and trustworthy confidence are separate properties, and that temperature scaling buys the second at no cost to the first; (iii) a robustness check against common image corruptions; (iv) a Grad-CAM analysis validated against random and centre-prior controls, which shows how much of an apparently good explanation is an artifact of dataset framing; and (v) a direct laboratory-to-field test on independently acquired imagery, which is where the headline accuracy stops holding.

II. RELATED WORK

A. Benchmark Performance and Laboratory-to-Field Generalisation

Plant diseases remain a major threat to agricultural productivity, food security, and farmers' livelihoods. The Food and Agriculture Organization of the United Nations estimates that plant pests and diseases account for losses of up to 40% of global crop production annually [1]. Consequently, automated plant-disease recognition has received considerable attention as a scalable alternative to manual inspection, which is time-consuming, subjective, and dependent on the availability of agricultural experts.

Early deep-learning studies demonstrated that convolutional neural networks (CNNs) can achieve near-saturated classification performance on curated plant-leaf datasets. Mohanty et al. [2] trained AlexNet and GoogLeNet models using 54,306 images from the PlantVillage dataset, covering 14 crop species and 38 crop-disease or healthy classes. Their best configuration achieved an accuracy of 99.35% on a held-out PlantVillage test set. However, when the same model was evaluated on independently collected images obtained from agricultural extension and online sources, its accuracy decreased to approximately 31%. This substantial reduction provided early evidence that high within-dataset accuracy does not necessarily translate into reliable performance under deployment conditions.

Ferentinos [3] extended this line of work to 87,848 images spanning 25 plant species and 58 plant-disease combinations; the best architecture reached 99.53% accuracy on unseen images drawn from the same distribution. That result confirmed the value of learned features for disease recognition, yet the imagery behind it remained highly standardised. Backgrounds are uniform, leaves sit centred in the frame, occlusion is rare, and illumination varies little. Such conditions make the task easier than it appears, and they leave room for a model to exploit cues specific to the acquisition setup rather than symptoms that transfer.

Domain shift accounts for most of the gap between laboratory and field performance. A photograph taken in a real plot rarely shows one clean leaf: soil, stems, weeds, shadows, hands and equipment all compete for the frame, leaves are partly occluded or seen at an angle, and symptoms appear at whatever stage the plant happens to be in. Camera quality, lighting, region, cultivar and season shift the visual distribution further. A random train-test split within one curated dataset cannot see any of this, so what it measures is within-distribution recognition rather than field generalisation.

Singh et al. [4] built PlantDoc to close some of that distance: 2,598 images across 13 plant species, photographed in the field rather than on a bench. The backgrounds are cluttered, the lighting inconsistent, the viewpoints varied and the occlusions natural. Their experiments showed that training on controlled images alone was not enough for real-world recognition, and that fine-tuning on PlantDoc cut classification error considerably. The dataset has since become a standard reference point for controlled-to-field transfer, though its modest size and skewed class distribution still make large-scale training difficult.

Recent surveys reach much the same conclusion about data. Kaya and Gürsoy [10] reviewed deep-learning pipelines for plant-disease classification, detection and segmentation, and named environmental variability, class imbalance, scarce field data and computational cost as the main obstacles to deployment. Nyawose et al. [11] report the familiar pattern of strong laboratory numbers falling away in the field, and trace it to training data that does not represent the deployment setting. Among the remedies they favour are multimodal learning, self-supervised pretraining, synthetic data, domain adaptation, and field datasets collected across more than one geography. The dependence of CNN-based image classification on acquisition conditions is not peculiar to agriculture; comparable transfer-learning pipelines applied to medical ultrasound imagery report the same sensitivity of learned features to how the images were captured [16].

Taken together, these studies suggest a simple reading. A score on a controlled benchmark is a first checkpoint, not proof of operational reliability. An evaluation aimed at deployment needs more than that. Testing across datasets, an explicit laboratory-to-field transfer, class-wise breakdowns, and some account of how the model behaves when background, lighting, occlusion or symptom appearance change all belong in it.

B. Lightweight Architectures and Deployment Efficiency

Accuracy is only half the problem. Large CNNs predict well but ask for compute and memory that a smartphone, an embedded processor, a drone or an Internet-of-Things node cannot always supply. Agricultural deployments are constrained on most axes at once, from processing power and storage through to battery, connectivity and acceptable latency, which is precisely why lightweight architectures have gained ground.

MobileNetV2 [5] introduced the inverted residual block with a linear bottleneck. Where a conventional residual block compresses and then expands, MobileNetV2 does the reverse, lifting a low-dimensional feature into a wider intermediate space, applying cheap depthwise convolutions there, and then projecting it back down. The pattern cuts parameters and multiply-accumulate operations sharply while keeping representational capacity competitive.

EfficientNet [6] takes a different route, scaling depth, width and input resolution together through a single compound coefficient. Growing one dimension in isolation tends to waste capacity; scaling all three in proportion spends the compute budget more evenly. The result is a family of variants that can be picked off the shelf to match a given accuracy, memory or latency target.

MobileNetV3 [7] combines hardware-aware architecture search with squeeze-and-excitation modules, retuned inverted residual blocks and cheaper nonlinearities. Because the search was run against mobile processors directly, the resulting network is a natural candidate for real-time agricultural use. The three designs are complementary rather than competing, and together they span a useful range of accuracy-versus-cost trade-offs.

These backbones transfer well in practice. Kalejahi et al. [13] compared MobileNetV2 with ResNet50 on roughly 87,000 plant-disease images and found MobileNetV2 reaching 98.33% accuracy with about seven times fewer parameters and a shorter training run, a strong argument for it in resource-constrained settings. The usual caveat applies. A single benchmark, or an augmented controlled dataset, says little about how the model will behave in real time in a real field.

Reporting efficiency well takes more than a parameter count. Model size, floating-point operations, memory footprint, latency, throughput and, where possible, energy draw all belong in the comparison. Latency in particular should be measured on the hardware the model will actually run on, because theoretical operation counts are a poor proxy; two networks with almost identical parameter counts can differ several-fold in wall-clock time, depending on memory access patterns, which operators the runtime supports, how well the work parallelizes, and what hardware acceleration is available.

Deployment rarely stops at local inference. A working monitoring system may tie together cameras, environmental sensors, edge processors, a wireless link and cloud analytics, and once data crosses those boundaries, secure transmission, confidentiality, access control, integrity and resilience all become part of the design. Mittal et al. [14] studied deep-learning analysis over secured data in a cloud-based sensor network, illustrating what security-aware processing involves when models sit inside connected sensing infrastructure. Security and privacy belong alongside predictive, computational and agronomic validation, however, as additional system-level requirements rather than substitutes for any of them.

C. Explainable Plant-Disease Recognition

Interpretability is particularly important in agricultural decision-support systems because an incorrect diagnosis may result in unnecessary pesticide application, delayed treatment, crop isolation, or financial loss. A model should therefore provide evidence that its predictions are based on biologically plausible symptoms, such as lesions, discoloration, necrosis, mildew, rust pustules, or deformation, rather than irrelevant background objects or image-acquisition artifacts.

Grad-CAM [8] produces class-discriminative localisation maps by weighting convolutional feature maps according to the gradients associated with a target class. The method can be applied to a trained CNN without modifying its architecture or retraining the model. It has therefore become one of the most widely adopted post-hoc explanation techniques for image-based plant-disease classification. Grad-CAM visualisations can reveal whether a model focuses on symptomatic leaf regions or relies on spurious cues such as background texture, image borders, shadows, or non-leaf objects.

Pal et al. [9] proposed Mob-Res, a lightweight architecture that combines MobileNetV2-based feature extraction with residual learning. The model contained approximately 3.51 million parameters and was evaluated on the PlantVillage and Plant Disease Expert datasets. In addition to classification performance and cross-domain validation, the authors applied Grad-CAM, Grad-CAM++, and Local Interpretable Model-Agnostic Explanations to visualise the regions influencing model predictions. Their study demonstrates that computational efficiency and post-hoc interpretability can be incorporated within a common plant-disease recognition framework.

Abrar and Kaur [12] surveyed where deep learning and explainable artificial intelligence meet in leaf-disease detection, covering CNNs, transfer-learning architectures, efficient models and newer hybrids, and arguing throughout for transparency and practical usability. Wider reviews of agricultural artificial intelligence reach the same position; explainability is not optional if the system is to be trusted [10], [11]. Comparable lightweight explainable frameworks outside agriculture report the same design conclusion, namely that interpretability can be retained under tight computational budgets rather than traded away for it [15].

Nevertheless, explanation analysis in plant-disease studies remains predominantly qualitative. Many studies present heatmaps for a small number of correctly classified examples and assess them through visual inspection. Such examples do not establish whether explanations are faithful, stable, class-specific, or consistently aligned with symptomatic regions. A visually plausible heatmap may also be produced for an incorrect prediction or may remain largely unchanged when the model output is perturbed.

Quantitative explanation evaluation can provide stronger evidence. Relevant measures include the proportion of attribution located within the leaf or lesion region, intersection-over-union with expert-annotated symptoms, explanation entropy, pointing-game accuracy, deletion and insertion scores, and stability under image perturbations. Explanation concentration is particularly useful for determining whether a model's attention is localised around disease symptoms or diffusely distributed across the image. Comparing explanation concentration between laboratory and field images can further reveal whether domain shift causes a model to redirect attention from pathological features toward environmental context. Explainability should not be interpreted as proof of causal or agronomic correctness. Post-hoc explanation methods describe the image regions associated with a model's output, but they do not independently validate the diagnosis. Their primary value lies in supporting model auditing, identifying spurious correlations, comparing architectures, and determining whether predictions are spatially aligned with expert expectations.

D. Research Gap Analysis

The existing literature establishes three major findings. First, CNNs can achieve very high accuracy on controlled plant-disease datasets [2], [3]. Second, architectures such as MobileNetV2, EfficientNet, and MobileNetV3 make resource-efficient deployment increasingly feasible [5]–[7], [13]. Third, explanation methods such as Grad-CAM can expose the spatial evidence associated with a model's predictions [8], [9], [12]. However, these dimensions are frequently investigated in isolation.

Many studies evaluate a single architecture, crop, or dataset and report overall accuracy as the principal measure of performance. Direct comparisons between lightweight models are often complicated by differences in input resolution, augmentation, class composition, optimisation strategy, transfer-learning protocol, and dataset partitioning. Moreover, field generalisation is frequently inferred from random test splits rather than evaluated through an independent field dataset.

Explainability is usually handled the same way, through a few illustrative heatmaps, with no quantitative check on localisation, concentration, stability or faithfulness. Rarely does a study ask whether two models with matching accuracy differ in how tightly their explanations sit on symptomatic tissue, and rarer still is any comparison of explanation behaviour between controlled and field images. Little published work therefore does all of the following at once:

1. compare several lightweight architectures under one consistent protocol;
2. evaluate multi-crop classification rather than a single crop–disease setting;
3. measure laboratory-to-field transfer directly, on independently acquired field imagery;
4. assess predictive performance and computational cost side by side; and
5. quantify how spatially concentrated the resulting explanations are.

Security and privacy are seldom weighed alongside accuracy, efficiency, transferability and explainability either, even though they matter as soon as a diagnostic model is embedded in a cloud-based sensor network [14]. That said, security is a system-level concern with its own literature, and we keep it analytically separate from the question of how well the recognition model itself performs. This study takes up each of those gaps directly, including laboratory-to-field transfer, which Section V-D measures on independently acquired field imagery and identifies as the necessary next step. We fine-tune three lightweight multi-crop plant-disease classifiers under one controlled protocol and judge them on three fronts at once: standard classification metrics with repeated seeds, computational cost, and a quantitative measure of explanation concentration tested against baselines. The aim is not simply to find the most accurate model, but to establish whether the accuracy differences between compact architectures are large enough to matter once they are measured properly.

III. METHODOLOGY

A. Datasets

Table I summarises the two datasets. PlantVillage [2] provides the laboratory imagery used for training and in-domain testing, covering 14 crops. Its photographs show a single leaf centred against a uniform background, and that framing matters both for the explanation results in Section V-C and for what the accuracy figures can be taken to mean. PlantDoc [4] provides the field test: in-the-wild photographs sharing 28 labels with PlantVillage once the differing taxonomies are mapped. It is used once, in Section V-D.

TABLE I
Datasets. PlantDoc class counts are after mapping onto the labels shared with PlantVillage; it is never trained on.

Dataset	Images / classes	Train / val / test	Purpose
PlantVillage [2]	54,305 / 38	38,013 / 8,146 / 8,146	Training, in-domain test
PlantDoc [4]	2,477 / 28	test only	Zero-shot field evaluation

B. Preprocessing

All images were indexed in a single manifest, and any class too small to survive stratification was dropped. What remained was split 70:15:15 by stratified sampling under a fixed random seed. The same manifest was reused for every run, so all reported numbers refer to one identical test set. Images were resized to 224×224 and normalised with ImageNet statistics; training samples additionally passed through random horizontal flips, rotations up to 25 degrees, and colour jitter.

C. Models and Transfer Learning

Three ImageNet-pretrained lightweight backbones were evaluated: MobileNetV3-Small [7], MobileNetV2 [5], and EfficientNet-B0 [6]. The original classifier in each network was replaced by a linear layer with 38 outputs, and the entire network was left trainable so that the convolutional features themselves adapt to leaf-disease appearance rather than remaining generic ImageNet features. Evaluating several candidate models under one identical protocol, rather than reporting a single architecture in isolation, follows established comparative-evaluation practice in which competing predictors are benchmarked side by side on a common dataset and metric set [17]. Fig. 1 summarises the framework and the experimental pipelines built on it.

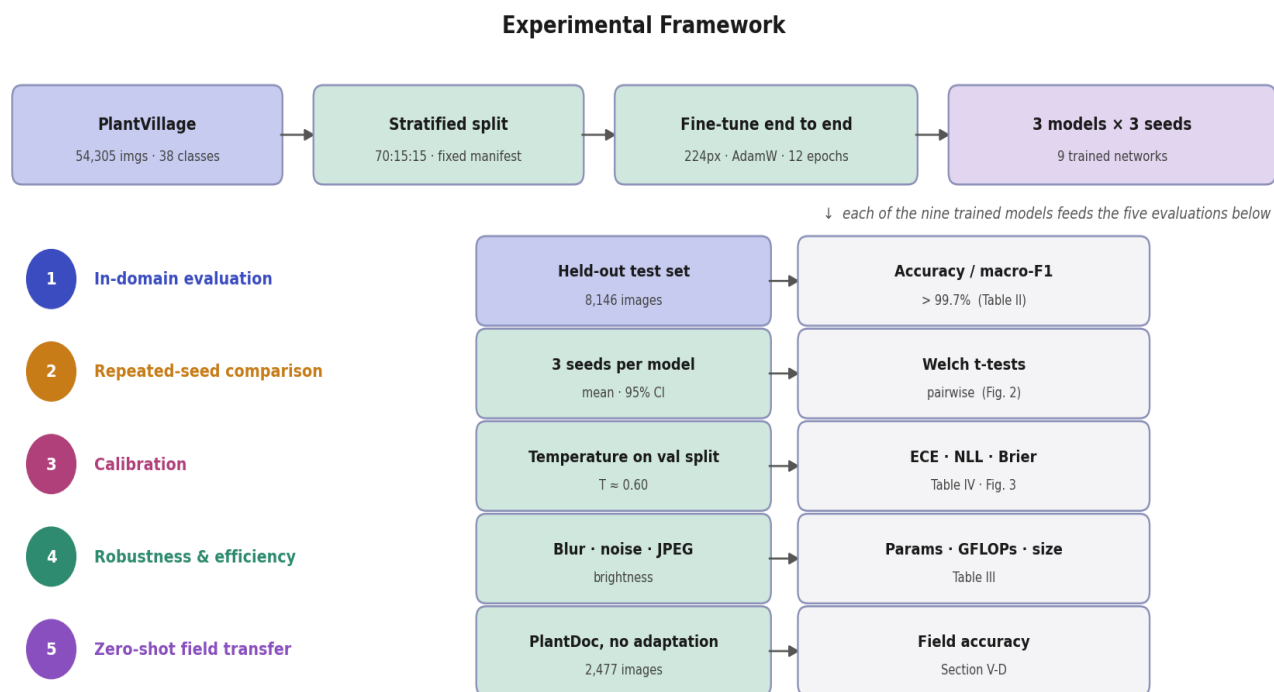


Fig. 1. Experimental framework: three backbones trained under three seeds each, then five evaluations.

D. Grad-CAM Explanation

For each prediction, Grad-CAM [8] was applied to the final convolutional layer. Spatially averaged gradients of the predicted-class score were used to weight the feature maps, and the rectified weighted sum formed the class-specific heatmap. Heatmap concentration was summarised with an energy-inside-leaf score. An approximate leaf mask was obtained from HSV saturation, and the score was computed as the Grad-CAM activation inside the mask divided by total activation.

E. Evaluation

Classification performance was measured using accuracy and macro-averaged precision, recall and F1 score; confusion matrices were used for error analysis. Because a single training run is a point estimate rather than a property of an architecture, every model was trained three times under different random seeds; we report the mean, its 95% confidence interval, and Welch's t-test between models, which does not assume equal variances. Accuracy alone says nothing about whether a model's confidence can be believed, so we additionally report expected calibration error over 15 bins, negative log-likelihood and the multiclass Brier score, before and after temperature scaling fitted on the validation split. Robustness was probed by re-evaluating the test set under four common corruptions—Gaussian blur, sensor noise, JPEG compression and a brightness shift. Efficiency was measured using trainable parameter count, multiply-accumulate cost as GFLOPs at the deployed resolution, and serialised model size. Explanation concentration was summarised by the mean energy-inside-leaf score against two controls, and generalisation by zero-shot transfer to PlantDoc.

IV. EXPERIMENTAL SETUP

Training was carried out on an NVIDIA RTX 3090 GPU. Images were processed at 224×224 with a batch size of 64. Every backbone was initialised from ImageNet weights and fine-tuned end to end for 12 epochs using AdamW with weight decay 1e-4, a cosine-annealing schedule over the full run, and mixed-precision arithmetic. The comparison in Section V-A used a learning rate of 3e-4 and 0.1 label smoothing; the calibrated runs in Section V-B used 2e-4 and 0.02, the lighter smoothing chosen because heavy smoothing distorts the confidence distribution that calibration is meant to correct. The checkpoint with the highest validation accuracy was retained. Each architecture was trained three times, under seeds 42, 7 and 2024, with the data split held fixed throughout so that seed variation reflects initialisation and batch ordering alone. These settings follow standard fine-tuning practice rather than a search: 224×224 is the resolution the ImageNet weights were trained at, 64 is the largest batch that fits the GPU at that resolution, and a learning rate an order of magnitude below the usual from-scratch value is the conventional adjustment when the weights already sit near a good solution. Twelve epochs was sufficient for validation accuracy to stop improving in every run. Temperature was fitted afterwards on the validation split by minimising negative log-likelihood, converging near 0.60 for all three seeds; a value below one indicates the uncalibrated networks were systematically underconfident.

V. RESULTS AND DISCUSSION

A. Classification Performance

Table II reports performance on the held-out PlantVillage test set for the best seed of each architecture. All three exceed 99.7% accuracy, with macro-F1 between 0.9970 and 0.9978. These figures sit above the 99.35% and 99.53% reported by Mohanty et al. [2] and Ferentinos [3] on comparable data, which confirms that a compact network, fine-tuned end to end, is not the limiting factor on this benchmark. Because one run is a point estimate, each model was retrained under three seeds; Table III reports the resulting means. Mean accuracies with 95% confidence intervals were 0.9974 [0.9967, 0.9982] for MobileNetV3-Small, 0.9980 [0.9974, 0.9985] for MobileNetV2, and 0.9984 [0.9981, 0.9988] for EfficientNet-B0; the standard deviations, between 0.0002 and 0.0003, indicate that run-to-run variation is far smaller than the gaps between architectures. Welch tests put EfficientNet-B0 significantly ahead of both MobileNetV3-Small ($p = 0.016$) and MobileNetV2 ($p = 0.040$), while MobileNetV3-Small and MobileNetV2 are statistically indistinguishable ($p = 0.075$). A single run per architecture would not have separated these models reliably, which is precisely why the repeated seeds matter. What the significance tests do not convey is how little is at stake: EfficientNet-B0's advantage over MobileNetV3-Small is 0.10 percentage points, roughly eight additional correct predictions out of 8,146, bought with 2.6 times the parameters, 6.8 times the floating-point cost and 2.6 times the model size (Table III). On a saturated benchmark, statistical significance and practical significance diverge. This is the reasoning behind cost-aware evaluation frameworks, which treat computational expenditure as a first-class evaluation axis alongside predictive quality rather than as a footnote to it [18]. We therefore read Table III, not Table II, as the operative comparison, and MobileNetV3-Small as the defensible deployment choice.

TABLE II
PlantVillage test-set classification performance (macro-averaged), best seed per model.

Model	Accuracy	Precision	Recall	Macro-F1
MobileNet V3-Small	0.9977	0.9968	0.9971	0.9970
MobileNet V2	0.9982	0.9968	0.9974	0.9971
EfficientNet-B0	0.9986	0.9976	0.9981	0.9978

TABLE III
Model efficiency at the deployed 224×224 input resolution.

Model	Parameters (M)	GFLOPs	Size (MB)	Mean accuracy (3 seeds)
MobileNetV3-Small	1.56	0.120	6.06	0.9974 ± 0.0003
MobileNetV2	2.27	0.638	8.90	0.9980 ± 0.0002
EfficientNet-B0	4.06	0.818	15.75	0.9984 ± 0.0002

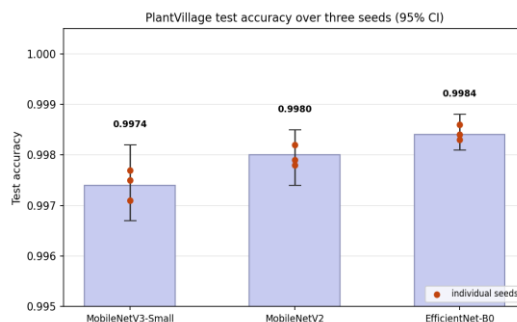


Fig. 2. PlantVillage test-set accuracy per model across three random seeds, with 95% confidence intervals.

B. Calibration and Robustness

High accuracy does not imply that a model's confidence can be believed, and for a diagnostic tool the second property matters nearly as much as the first. Table IV reports both. Straight out of training the networks are accurate but poorly calibrated, with an expected calibration error of 0.0193. They are systematically under-confident: mean confidence is about 97.9% on predictions that are right 99.85% of the time, and the fitted temperature near 0.60 reflects this. Temperature scaling is a single scalar fitted on the validation split, so it cannot change any prediction and leaves accuracy and macro-F1 untouched, yet it reduces calibration error to 0.0009 and negative log-likelihood from 0.0250 to 0.0077. Averaging the three calibrated seeds improves matters slightly again, reaching 0.9986 accuracy with a calibration error of 0.0005 and a bootstrap 95% interval of [0.9978, 0.9994]—11 errors in 8,146 images.

Robustness was probed by re-evaluating the calibrated ensemble under four corruptions. The losses are slight. A brightness shift to 65% costs 0.02 percentage points, JPEG compression at quality 35 costs 0.06, Gaussian blur 0.11, and sensor noise at ± 20 levels 0.14. Calibration degrades proportionally, the error rising from 0.0005 to at most 0.0018, so the confidence estimates remain usable under mild image degradation. Fig. 3 shows the effect directly: before scaling, accuracy runs consistently above confidence; afterwards the two coincide across the range that carries the data. Neither addition costs anything at deployment. Temperature scaling divides the logits by one stored scalar before the softmax, leaving inference time unchanged, and the corruption study is an offline measurement that adds nothing to the served model.

TABLE IV
Calibration on the PlantVillage test set. The first two rows are means over three seeds.

Setting	Accuracy	Macro-F1	ECE	NLL
Uncalibrated	0.9985	0.9980	0.0193	0.0250
Calibrated	0.9985	0.9980	0.0009	0.0077
3-seed ensemble	0.9986	0.9981	0.0005	0.0063

Reliability before and after temperature scaling

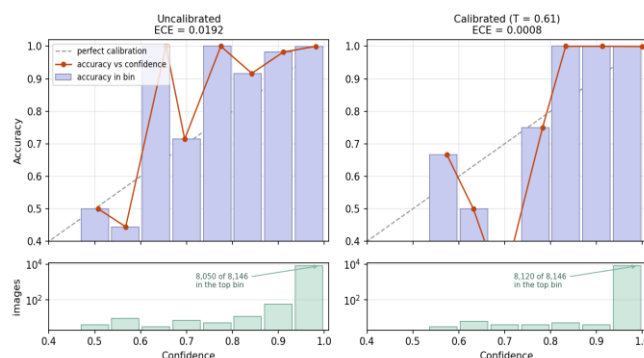


Fig. 3. Reliability before and after temperature scaling. The lower panels give the confidence distribution; almost every prediction falls in the highest-confidence bin, so the sparse low-confidence bins carry only a handful of images.

C. Explainability

Qualitative inspection indicated that Grad-CAM activation was concentrated on symptomatic leaf regions for most correct predictions and became more diffuse for errors. Across 120 test images the mean energy-inside-leaf score was 0.513 ± 0.153 . A concentration figure means little on its own, so it was compared against two controls on the same images. Grad-CAM sits far above a random heatmap (0.379 ± 0.134 ; paired $t = 16.74$, $p < 0.001$), confirming that the attention is not accidental. A plain centre-weighted Gaussian, however, scores 0.548 ± 0.175 and remains significantly above Grad-CAM (paired $t = -4.53$, $p < 0.001$). PlantVillage leaves are centred in the frame, so a large part of the apparent localisation is explained by that framing rather than by lesion-specific evidence. That the centre prior stays ahead even for networks whose features were tuned specifically on leaf disease suggests the bias belongs to how the dataset was photographed rather than to what the model learned. We therefore treat the score as evidence that the model attends to the leaf, not as evidence that it has localised the lesion.

D. Transfer to Field Imagery

The final experiment measures how far the models carry to photographs taken in the field. Each was applied to PlantDoc [4] without adaptation, using 2,477 images from real plots mapped onto the 28 labels the two datasets share. Accuracy on this set is MobileNetV2 20.0%, MobileNetV3-Small 15.1%, EfficientNet-B0 13.5%. Restricting the prediction to the overlapping labels raises this to 17.4–22.6%. Chance over 28 classes is 3.6%, so the models retain real signal on field imagery, well below their in-domain accuracy.

Two observations follow. Fine-tuning helps here as well as in-domain. These networks roughly double the field accuracy of the frozen-feature versions, so stronger features do transfer. What they do not overcome is the difference in the photographs themselves. Field images carry cluttered backgrounds, variable daylight, several leaves at once, and symptoms at stages the curated set does not cover. The ordering inverts completely. EfficientNet-B0, the most accurate model in-domain, is the least accurate in the field, while MobileNetV2, placed second on PlantVillage, is the strongest here. In-domain accuracy is therefore a weak guide to field behaviour, and this gap is best read as the target for domain adaptation rather than as a property of the architectures.

E. Comparison with Prior Work

Table V places these results beside representative prior work. Mohanty et al. [2] and Ferentinos [3] established that curated benchmarks are close to saturated, at 99.35% and 99.53%, and the accuracy reported here is comparable while using a network small enough to run on a phone. Among recent lightweight studies, Pal et al. [9] pair MobileNetV2 features with residual learning at 3.51 M parameters, and Kalejahi et al. [13] report 98.33% with MobileNetV2. Against these, the present work is the most economical of the group at 1.56 M parameters and 0.120 GFLOPs, and it adds three things the comparison set does not report: calibrated confidence in place of raw softmax scores, a robustness check under common image corruptions, and a zero-shot field measurement. The limitation it shares with them is field accuracy. The drop on PlantDoc follows from the training data rather than from any one architecture, and it remains open across this line of work.

TABLE V
Comparison with prior work; a dash marks an unreported quantity.

Study	Acc.	Params	ECE	Field
Mohanty [2]	99.35%	—	—	31%
Ferentinos [3]	99.53%	—	—	—
Pal et al. [9]	—	3.51	—	Yes
Kalejahi [13]	98.33%	2.27	—	—
This work	99.74%	1.56	0.0005	15.1%

F. Limitations

Two caveats bound the scope of these results. The field evaluation is zero-shot, so it establishes the size of the gap but not how much of it domain adaptation would close; calibration is likewise fitted in-domain and would need refitting after adaptation. The Grad-CAM analysis uses a heuristic HSV leaf mask and is validated against random and centre-prior controls rather than expert-annotated lesions, which neither dataset provides.

VI. CONCLUSION AND FUTURE WORK

This study examined how far a compact, explainable pipeline can be taken for multi-crop leaf-disease classification, and where it stops. Fine-tuned end to end, all three lightweight backbones exceed 99.7% accuracy on PlantVillage, above the published baselines for that benchmark, and the repeated-seed study shows this is stable rather than fortunate. Temperature scaling makes the confidence trustworthy as well as the prediction accurate, cutting expected calibration error by a factor of twenty at no cost to accuracy, and the models absorb blur, noise, compression and brightness shifts almost without loss. Two findings carry the most practical weight. The accuracy differences between these architectures are statistically detectable but small, a tenth of a percentage point against several-fold differences in cost, so on a saturated benchmark efficiency should decide the deployment choice; MobileNetV3-Small is the strongest candidate on that basis. The zero-shot field measurement then sets the agenda for the next stage of the work: domain adaptation on field imagery, with recalibration afterwards rather than an assumption that calibration survives the move.

REFERENCES

- [1] Food and Agriculture Organization of the United Nations, "Plant production and protection." [Online]. Available: <https://www.fao.org/plant-production-protection/about/en>. Accessed: Jul. 19, 2026.
- [2] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, art. no. 1419, 2016, doi: 10.3389/fpls.2016.01419.
- [3] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 2018, doi: 10.1016/j.compag.2018.01.009.
- [4] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, "PlantDoc: A dataset for visual plant disease detection," in *Proc. 7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD)*, Hyderabad, India, 2020, pp. 249–253, doi: 10.1145/3371158.3371196.
- [5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.
- [6] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 6105–6114.
- [7] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

- [8] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
- [9] C. Pal, S. Karmakar, I. Mukherjee, and P. P. Chakrabarti, “A lightweight and explainable CNN model for empowering plant disease diagnosis,” *Scientific Reports*, vol. 15, art. no. 30720, 2025, doi: 10.1038/s41598-025-94083-1.
- [10] Y. Kaya and E. Gürsoy, “A review of deep learning architectures for plant disease detection,” *Turkish Journal of Biology*, vol. 49, no. 5, pp. 459–497, 2025, doi: 10.55730/1300-0152.2761.
- [11] T. Nyawose, R. C. Maswanganyi, and P. Khumalo, “A review on the detection of plant disease using machine learning and deep learning approaches,” *Journal of Imaging*, vol. 11, no. 10, art. no. 326, 2025, doi: 10.3390/jimaging11100326.
- [12] O. Abrar and S. Kaur, “Deep learning meets explainable AI: A review for plant leaf disease detection,” in Proc. 2025 International Conference on Intelligent and Innovative Practices in Engineering and Management (IIPeM), 2025, doi: 10.1109/IIPeM65914.2025.11547951.
- [13] B. K. Kalejahi, S. Yahyayeve, I. Abdurahmanova, and S. Hasanzade, “Lightweight deep learning models for early plant disease detection: A transfer learning approach with MobileNetV2 and ResNet50,” *Problems of Information Society*, vol. 17, no. 1, pp. 11–23, 2026, doi: 10.25045/jpis.v17.i1.02.
- [14] H. K. Mittal, M. Sharma, L. Sehgal, and V. Jain, “Analysis of highly secured data through deep learning for cloud-based sensor network,” in Proc. 2026 Eighth International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonapat, Haryana, India, 2026, pp. 792–798, doi: 10.23919/CCICTCPS00010.2026.00130.
- [15] S. Singla, H. K. Mittal, S. S. Kang, P. Singla, C. Senthilpari, and R. Azmeera, “Lightweight explainable AI framework for early academic risk prediction and sustainable student support in engineering education,” *Social Science Forum*, vol. 10, no. 7, pp. 655–675, 2026.
- [16] S. M. Harun Or Rashid, H. Mittal, M. Thompson, A. A. Hamid, S. Assi, and D. A.-J. Obe, “Breast cancer prediction from breast ultrasound images using deep learning,” in Proc. 2026 International Conference on Electrical/Electronics, Robotics, Artificial Intelligence, and Informatics (ICERAI), 2026, doi: 10.1109/ICERAI69511.2026.11494592.
- [17] H. K. Mittal, P. Dalal, P. Garg, and R. Joon, “Forecasting pollution trends: Comparing linear, logistic regression, and neural networks,” in Proc. 2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP), 2024, pp. 411–419, doi: 10.1109/INNOCOMP63224.2024.00074.
- [18] H. K. Mittal, “Cost-aware evidence-triage routing for biomedical RAG: A leakage-controlled evaluation framework,” *Social Science Forum*, vol. 10, no. 7, pp. 508–529, 2026.