Elinvo (Electronics, Informatics, and Vocational Education)
Vol 11, No. 1, 2026, pp. 68 - 84
ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

Data-Centric Preprocessing Outperforms Loss Modifications for Hard-Class Plant Disease Classification Using MobileNetV3

Bambang Priambodo[1*,2], Abdul Fadlil[3], Sunardi[4]

[1*]Doctoral Program in Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, 55191
[2] Doctoral Program in Informatics, Universitas Saintek Muhammadiyah, Jakarta, Indonesia 13730
[3,4] Doctoral Program in Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia 55191
[1*]2436083026@webmail.uad.ac.id; [3] abdul.fadlil@uad.ac.id; [4]sunardi@uad.ac.id
*corresponding author

Article Info

Article history:
Receive, April 27, 2026
Review, June 14, 2026
Accepted, June 23, 2026
Published, July 16, 2026

Keywords:
data-centric pipeline; image quality assessment; plant disease classification; targeted augmentation; lightweight CNN

Abstract

Data-centric approaches have gained attention in plant disease classification; however, a systematic evaluation of underperforming 'hard classes' remains limited. This study proposes a four phase pipeline comprising granular error diagnosis, no-reference image quality assessment, class-targeted augmentation, and an ablation study to improve hard-class robustness without increasing model complexity. Using the New Plant Disease dataset and a lightweight MobileNetV3-Small backbone, we first established a baseline. Based on this baseline performance, we identified 12 hard classes (defined as those with either F1 < 0.96 or recall < 0.96 on the baseline model) with a mean F1 of 0.9260. The optimal configuration (categorical cross-entropy, no class weighting, baseline head) raised the mean hard-class F1 to 0.9668, corresponding to an absolute improvement of +4.08 percentage points, while maintaining global test accuracy at 98.08%. A two-tier design with three random seeds confirmed robustness (mean hard-class F1 = 0.9652 ± 0.0020). Unexpectedly, after data enrichment, inverse frequency class weighting degraded hard-class F1 (to 0.9621), and the default focal loss parameters offered no additional benefit over plain cross-entropy (0.9647). The MobileNetV3-Small showed comparable performance to the heavier EfficientNetB0 under our evaluation protocol, with no statistically significant difference detected (p = 0.089; limited power due to n = 3). Tomato Target Spot (C35) remained the most persistent bottleneck (F1 ≈ 0.90). Future work includes explainable AI for error analysis and real-field validation.

INTRODUCTION

Lightweight CNN backbones such as MobileNetV3 and EfficientNet are now the standard for edge-based plant disease detection, balancing computational efficiency with inference performance [1], [2]. Later works have improved representation via multi-stage feature extraction, dynamic dilated convolution, local attention, and squeeze-and-excitation mechanisms without significantly increasing parameters [3], [4], [5].

Following this trend, Verma et al. [6] proposed ViTCon, a hybrid CNN–ViT model using self-attention to capture long-range spatial dependencies, achieving up to 99.56% accuracy and near-99.5% F1 on public corn, rice, and wheat disease datasets (including PlantVillage-based data). However, ViT-based and hybrid models demand far greater computational resources than lightweight CNNs (ViT-L32 has approximately 303 million parameters compared to a few million for MobileNet-family models) posing deployment barriers for resource-constrained edge devices [7]. Furthermore, these architectures

are mostly evaluated with global accuracy and macro-F1, with limited systematic analysis of per-class robustness, especially for underperforming hard classes.

With the rise of data-centric learning, studies have shown that improving data quality through targeted augmentation, label correction, and balancing can outperform model-centric approaches without changing network architecture [8], [9], [10], [11]. Meanwhile, Image Quality Assessment (IQA) has been used to filter low-quality images in computer vision and medical imaging [12], [13]. However, in plant disease detection, data-centric and IQA strategies remain partial and global—not selectively targeted at bottleneck or hard classes. Notably, no-reference IQA methods such as Laplacian variance and entropy measures are well studied for generic and medical images but have not been systematically adopted as automated preprocessing filters in plant leaf disease classification. According to recent surveys, IQA filtering based on Laplacian variance or entropy is not a standard practice in this domain [14], [15].

The problem of class imbalance is typically resolved through adjustments to the loss function (e.g., focal loss, class weighting) [16], [17], [18] and is evaluated using global metrics such as accuracy and macro-F1 [19]. Some studies have started to report per-class analysis using confusion matrices [20], [21] but a specific evaluation of hard classes, e.g., the mean F1 score, has not been a primary metric for success, especially in realistic and less balanced data distributions.

A comprehensive review of recent studies on leaf disease classification summarised numerous studies on the use of deep learning architectures and data augmentation on the PlantVillage (PV) dataset [14]. In general, the reviewed studies reported high global accuracy using overall F1 scores and confusion matrices; however, none treated weak or hard classes as a systematic evaluation target. For example, Ashurov et al. and Imanulloh et al. [5], [22] reported global metrics of 97-98% across 38 classes using the PV and New Plant Disease (NPD) datasets. Although both studies presented a confusion matrix (CM), only the study by Ashurov et al. provided a partial analysis of the underperformance of specific classes, explicitly acknowledging visual evidence of such weaknesses. To illustrate this trend more clearly, we summarise representative studies on PV and NPD in Table 1. From this comparison, we identify three common limitations that require further attention.

Table 1. Summary of related works and identified research gaps

Authors (years)	Architecture	Dataset (classes)	Hard-class Eval.	Per-class Eval.	IQA	Targeted Augmentation	Accuracy
Ashurov et al.	Modified MobileNetV2	PV (38)	Partial	Yes	No	No	98.0%
Priambodo et al. (2025)	MobileNetV3-Small, EfficientNetB0	NPD (38)	Partial	Yes	No	No	97-98%
Siraj & Ansari (2025)	ParaleafNet	PV (38)	No	Partial (CM)	No	No	99.56%
Ullah et al. (2023)	DeepPlantNet	PV (8)	Partial	Yes	No	No	98.49%
Guan et al. (2023)	Dise-Efficient	PV (38)	No	No	No	No	99.80%
Imanulloh et al. (2023)	Custom CNN	NPD (38)	No	Partial (CM)	No	No	97-98%
Sun et al. (2022)	EfficientNet + focal loss ResNet50	NPD (10)	Partial	Partial (metric table)	No	No	99.7% 95.1%
Hassan et al. (2021)	InceptionV3, MobileNetV2	PV (38)	No	No	No	No	98.42%

From Table 1, we identify three common weaknesses that we believe require further attention, as overlooking these limitations may lead to an overly optimistic interpretation of model performance:

- **No explicit hard-class evaluation**. Some studies do report per-class F1 scores or confusion matrices, but they still prioritise global accuracy and macro-F1 as their main performance measures. To the best of our knowledge, none of these studies pre-define a set of "hard" or bottleneck classes and track an aggregate metric—for example, the mean F1 of such weak classes—as a key indicator of success. This practice risks obscuring systematic weaknesses in specific classes that are crucial for real-world deployment.
- **No IQA as a data filter**. Laplacian variance and Shannon entropy have not been systematically integrated into plant disease classification pipelines as a pre-training data governance step.
- **No class-targeted augmentation**. Data augmentation, when applied, is typically uniform across all classes. Augmentation that is selectively intensified only on underperforming classes—guided by empirical performance diagnosis—is absent.

Moreover, a closer inspection of the NPD dataset reveals that the original "test" set contains only 8 out of 38 classes with 2–6 unlabelled images each—insufficient for rigorous evaluation. Consequently, previous studies have either used ad-hoc splits or relied solely on validation accuracy, further obscuring class-level generalisation.

We define a hard class as any class with either F1 < 0.96 or recall < 0.96 (or both) on the baseline MobileNetV3-Small model. This dual-threshold criterion was chosen for two reasons. First, a clear gap consistently appears in the per-class F1 distribution across three independent runs, with 26 classes achieving both F1 ≥ 0.96 and recall ≥ 0.96, while 12 classes fall below at least one of these thresholds. Second, in agricultural diagnostics, an F1 below 0.96 indicates suboptimal precision-recall balance, while a recall below 0.96 implies missing more than 4% of infections—both of which are operationally detrimental. We therefore use the combined F1 and recall criteria as a comprehensive indicator of per-class diagnostic quality. Based on this definition, 12 hard classes were identified from the baseline performance.

Given the identified gaps, this study addresses the following research questions:

- **RQ1**. To what extent can an IQA-guided, class-targeted augmentation pipeline improve the F1 scores of the hard classes for MobileNetV3-Small on the NPD dataset, while preserving global performance?
- **RQ2**. After improving data quality and class representation through IQA and targeted augmentation, do conventional imbalance mitigation strategies (inverse-frequency class weighting and focal loss) still provide measurable benefit, or do they become redundant or detrimental?
- **RQ3**. How do different classifier head configurations (baseline, deeper, lighter) affect the trade-off between global accuracy and hard-class robustness?

This paper makes three main contributions:

- **Hard-class-centric evaluation protocol**. We define hard classes as baseline classes with either F1 < 0.96 or recall < 0.96, and use the mean F1 of these classes as a primary success metric alongside global accuracy and macro-F1S. The mean F1 computed over these hard classes then becomes the main success metric, together with global accuracy and macro-F1.
- **Data-centric pipeline**. We propose a four-stage pipeline that combines no-reference IQA (Laplacian variance and Shannon entropy based) and class-targeted augmentation to improve hard-class performance on MobileNetV3-Small.
- **Counterintuitive empirical evidence**. We show that the simplest setup, plain categorical cross-entropy with a baseline head and no class weighting, yields the best hard-class F1 even when

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

using IQA and targeted augmentation, outperforming both focal loss and inverse-frequency class weighting.

The rest of the paper is organised as follows. Section 2 describes the proposed methodology. Section 3 describes experimental setup, implementation details, results and discussion. In Section 4 we conclude the paper and outline directions for future work.

METHODS

Before model training, we prioritised diagnosis of data quality rather than modifying the model architecture. The phased experimental framework is summarised in Figure 1, which also shows the interaction of the four phases in identifying and addressing these hard classes. The framework includes four phases: (1) granular diagnosis and error profiling, (2) IQA pipeline, (3) targeted data augmentation strategy, and (4) an ablation study.

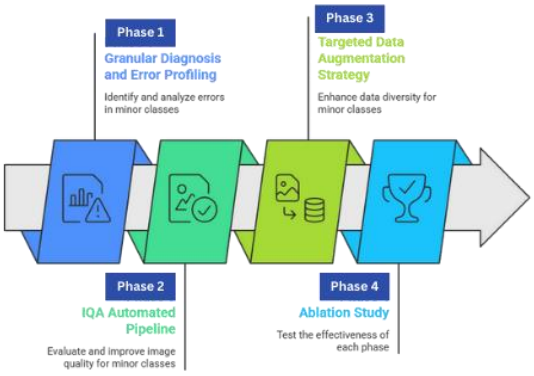

Figure 1. Four phase research methodology

Phase 1: Granular Diagnosis and Error Profiling

To expose class-specific weaknesses, we perform fine-grained error profiling at the class level, by quantifying each class individually, and thus avoiding the pitfalls of global metrics that can hide such disparities [23], [24]. The dataset we chose is the NPD dataset from Kaggle (https://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset), which is a curated version of the original PV dataset [25]. For the data, the creators used augmentations such as rotation, flipping, zooming, changing brightness, and shifting, and they also down sampled the data to reduce imbalance across classes. The original PV (2015) contains approximately 53,000 images, whereas NPD provides 85,486 RGB images of healthy and diseased leaves of 38 classes from 14 plant species.

The NPD dataset has been widely used as a benchmark in the studies of plant leaf disease classification [22], [23], [26] due to its better class balance than the original PV dataset. The comparison of the class distribution between PV and NPD is illustrated in Figure 2, and detailed class codes and image counts are summarised in Table 2.

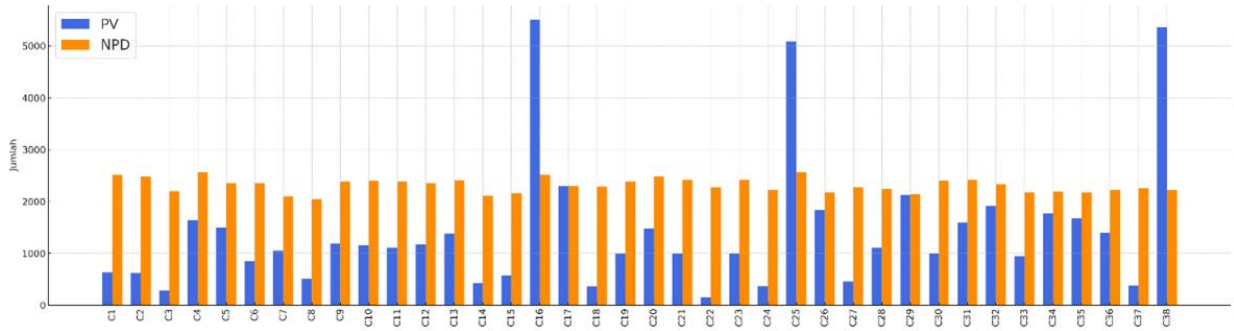

Figure 2. Class distribution comparison between PV and NPD

The original NPD split provides sufficient training and validation data (approximately 2,000-2,500 images per class), but the provided test set only has 8 of the 38 classes with 2-6 unlabelled images per class, which is insufficient for a proper evaluation. Therefore, a dedicated test set was constructed by randomly withholding 15% of samples per class from the training data via stratified sampling (seed = 1337) while retaining the original validation set. This yielded a 68/20/12 train/validation/test split with all 38 classes represented (approximately 250–300 test images per class). All images were resized to 224 × 224 pixels.

Table 2. Overview of NPD dataset classes grouped by plant species

Class Name	Class Code	Qty. (Train+Valid)	Class Name	Class Code	Qty. (Train+Valid)
Apple scab	C1	2,520	Pepper bell healthy	C20	2,485
Apple black rot	C2	2,484	Potato early blight	C21	2,424
Apple cedar rust	C3	2,200	Potato late blight	C22	2,280
Apple healthy	C4	2,510	Potato healthy	C23	2,424
Blueberry healthy	C5	2,270	Raspberry healthy	C24	2,226
Cherry (including sour) powdery mildew	C6	2,104	Soybean healthy	C25	2,557
Cherry (including sour) healthy	C7	2,352	Squash powdery mildew	C26	2,170
Corn (maize) cercospora leaf spot gray leaf spot	C8	2,052	Strawberry leaf scorch	C27	2,280
Corn (maize) common rust	C9	2,384	Strawberry healthy	C28	2,219
Corn (maize) northern leaf blight	C10	2,389	Tomato bacterial spot	C29	2,137
Corn (maize) healthy	C11	2,324	Tomato early blight	C30	2,400
Grape black rot	C12	2,360	Tomato late blight	C31	2,416
Grape esca (black measles)	C13	2,410	Tomato leaf mold	C32	2,352
Grape leaf blight (isariopsis leaf spot)	C14	2,152	Tomato septoria leaf spot	C33	2,181
Grape healthy	C15	2,115	Tomato spider mites two-spotted spider mite	C34	2,176
Orange huanglongbing (citrus greening)	C16	2,513	Tomato target spot	C35	2,284
Peach bacterial spot	C17	2,297	Tomato yellow leaf curl virus	C36	2,451
Peach healthy	C18	2,165	Tomato mosaic virus	C37	2,238
Pepper bell bacterial spot	C19	2,391	Tomato healthy	C38	2,314

A baseline MobileNetV3-Small was trained on the original training set using categorical cross-entropy and a standard classifier head (GlobalAveragePooling2D → Dense256 → Dropout0.5 → Dense38). Training followed a two-stage transfer learning protocol: 5 epochs with a frozen backbone (lr = 1e-3), then 30 epochs fine-tuning the last 20 layers (lr = 1e-5). Early stopping (patience = 5) and ReduceLROnPlateau (factor = 0.5, patience = 3) were also applied.

Table 3. Initial performance of MobileNetV3-Small on the original NPD dataset

Metric	Value
Test accuracy	0.9670
Macro F1	0.9667
Weighted F1	0.9667
Mean F1 (12 hard classes)	0.9260

Per-class precision, recall, and F1 were evaluated on the validation set, revealing a natural bimodal split: 26 classes attained F1 ≥ 0.96 (mean = 0.990), whereas the remaining 12 classes scored ≤ 0.96 (mean = 0.926), with no class in the [0.94, 0.96] interval. The baseline per-class F1 and recall histograms, illustrated in Figure 3, reveal a pronounced bimodal pattern. Eleven classes scored below 0.96 in F1, while the remaining 27 clustered above 0.97, leaving a visible gap around 0.96 (Figure 3a).

The recall distribution in Figure 3b shows a similar trend. We included Class C12, which despite having an F1 score of 0.971, had a recall of 0.951. This decision was based on a risk-aware framework, adapted by placing greater emphasis on false negatives [27] and adjusting the classification cost while maintaining balance (Table 4). The confusion matrix as shown in Figure 4 highlights common misclassifications, such as C8→C10 (27 images) and C30→C31 (16 images), reinforcing the need for targeted augmentation [3], [23]. Global baseline metrics are displayed in Table 3, and Table 4 lists the 12 hard classes.

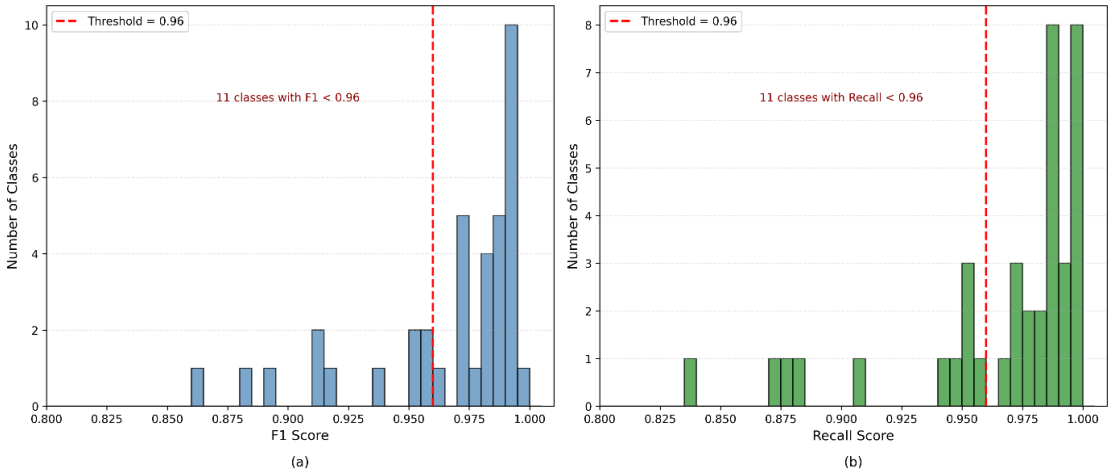

Figure 3. Histogram of per-class (a) F1 scores and (b) recall scores for the baseline MobileNetV3-Small model on the NPD

Table 4. 12 hard classes identified from baseline performance

Code	F1	Recall
C8	0.9118	0.8785
C10	0.9396	0.9756
C12	0.9712	0.9507
C22	0.9519	0.9519
C23	0.9573	0.9416
C29	0.9515	0.9533
C30	0.8623	0.8368
C31	0.8936	0.9065
C32	0.9593	0.9576
C33	0.9149	0.8817
C34	0.9168	0.9466
C35	0.8824	0.8727

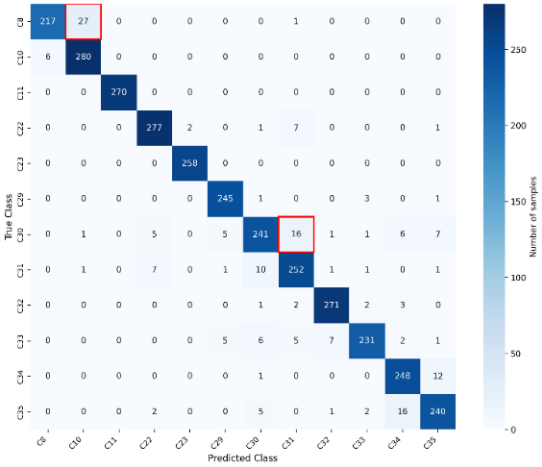

Figure 4. Confusion matrix for the 12 hard classes

Phase 2: IQA Pipeline

In Phase 2, selective quality filtering removed low-quality samples associated with hard-class errors by using an automated no-reference IQA mechanism for objective and reproducible data cleaning [12]. This phase comprised blur detection and information-richness evaluation.

In the blur detection stage, image sharpness is quantified using the variance of the Laplacian (LAPV), a no-reference focus measure that captures second-order intensity variations related to edges and fine texture [28], [29]. Let $I(x, y)$ denote the greyscale image and $L(x, y)$ the response of a discrete Laplacian operator applied to $I(x, y)$. The global Laplacian variance ϕ for an image is computed as:

$$\emptyset = \frac{1}{N}\sum_{x,y}(L(x,y) - \bar{L})^2 \quad (1)$$

$\bar{L}$ is the mean Laplacian response over all N pixels. Lower values of ϕ indicate weak edge structure and thus stronger blur [29]. Images with ϕ below an empirical threshold ($\phi < 150$ in our experiments) are classified as blurry and removed from the training dataset to prevent the model from learning ambiguous features. The global LV is computed over all pixels of the image, including the dark background regions that are typical of NPD leaf images. To mitigate the risk that these largely homogeneous backgrounds artificially depress the focus measure, the blur threshold (LV = 150) was chosen conservatively via sensitivity analysis so that only severely blurred samples are removed while the overall class-wise training distribution remains stable.

In the Shannon Entropy (SE) evaluation, the complexity of visual information is quantified using traditional entropy [30], calculated as:

$$H(x) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \quad (2)$$

In image analysis, SE $H(x)$ quantifies visual complexity from the intensity distribution of pixels, where x_i denotes the i-th discrete intensity level, $P(x_i)$ its probability (normalised histogram), and n the total number of levels; the term $\log_2 P(x_i)$ measures the information per level, so $H(x)$ reflects intensity diversity. We discarded the images with low-entropy because glare or dominant backgrounds can obscure useful details, such as leaf texture [30]. In contrast, images with high-entropy exhibit richer textures and more informative features.

We analysed how LV and SE were distributed across the 12 hard classes. The 10th percentiles, LV = 479 and SE = 6.4, marked the point at which detectable quality degradation began. Based on the sensitivity analysis results (Table 5), an LV of 150 eliminated approximately 11–13% of images while preserving class balance [13]. When the LV threshold varied from 100 to 200, most classes remained stable; however, class C31 underwent a significant shift from 27.1% to 29.4%. Class C35 showed an average change of 8.8%, indicating moderate sensitivity, whereas C12, C22, C30, and C34 remained unchanged.

Table 5. Summary of sensitivity analysis for IQA thresholds

Parameter	Tested range	Threshold Parameters	Global elimination rate	Sensitive class(es)	Stable classes (0% elimination)
LV	100 – 200	150	11.12% – 12.83%	C31 (27.1% – 29.4%); C6 (18.9% – 42.9%); C16 (35.7% – 41.5%); C36 (28.5% – 43.5%)	C12, C13, C14, C21, C27, C34
SE	5.5 – 7.0	6.4	11.98% (at SE=6.4)	C8, C10, C29, C32, C33 (sharp jumps > 6.4); also C9, C26, C6	C12 (all levels); C34 (≤ SE=6.4)

With SE values of 6.4, all classes exhibited failure rates below 10%. Increasing SE from 6.4 to 6.8 caused substantial increases across multiple classes: C8 (4.6% to 33.3%), C10 (3.6% to 41.5%), C29 (4.9% to 30.4%), C32 (6.8% to 34.6%), and C33 (8.4% to 30.6%), indicating over-filtering. At the low end (SE = 5.5), global removal was negligible (0.1%), while at SE = 7.0 it climbed steeply (C8: 52.4%,

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

C10: 71.3%, C29: 75.4%). C12 remained stable and C34 rose to 1.7% at SE = 6.8. We assessed that C22 and C30 increased above the threshold value, therefore we identified them as being unstable.

The global elimination rate was calculated to be 11.98% based on LV = 150 and SE = 6.4, and this figure appears to be a reasonable estimate. Table 6 displays the elimination and retention rates for the 12 hard classes. The elimination rates, ranging from 0.0% (C12, C22, C29, and C30) to 28.2% (C31), determine the available subset for further augmentation. In an effort to avoid data leakage, we implemented filters only on the training set, while verification and evaluation sets remained unchanged.

Table 6. IQA elimination rates for the 12 hard classes (LV = 150, SE = 6.4)

Class	Original training samples	IQA-passed samples	Elimination rate (%)
C8	1,395	1,331	4.6
C10	1,625	1,576	3.6
C12	1,604	1,604	0.0
C22	1,648	1,641	0.4
C23	1,550	1,440	7.1
C29	1,455	1,384	4.9
C30	1,632	1,626	0.4
C31	1,573	1,130	28.2
C32	1,599	1,491	6.8
C33	1,483	1,358	8.4
C34	1,479	1,479	0.0
C35	1,552	1,416	8.8

Phase 3: Targeted Data Augmentation Strategy

After the diagnostic phase, we performed augmentation only on the hard classes and the IQA-passed training set, consistent with the previous research [9], [10], [11]. By using Python Imaging Library (PIL), we performed image augmentation with five variants per image using methods including random rotation (−25° to 25°), horizontal flips, colour jitter (brightness and contrast: 0.7–1.3), and Gaussian noise (0.5–1.5). This strategy was designed to improve the representation of features near weak decision boundaries and to increase resilience to pose and lighting variations, which are common in practical deployment. We preserved the dataset's natural imbalance to support a more realistic evaluation.

Phase 4: Ablation Study

We conducted a systematic ablation study with seven scenarios as shown in Table 7 to identify the contribution of each component and identify the optimal combination that maximises hard-class F1 without sacrificing global performance. Table 7 compares two imbalance strategies: loss function modification and class weighting.

Table 7. Ablation study scenario configuration

Scenario	Loss Function	Class Weighting	Classifier Head
S0	CCE	N	Baseline
S1	CCE	Y (Inv. Freq)	Baseline
S2	CCE	Y (Inv. Freq)	Deeper
S3	FL ($\gamma = 2$, $\alpha = 0.25$)	Y (Inv. Freq)	Baseline
S4	CCE	Y (Inv. Freq)	Lighter
S5	FL ($\gamma = 2$, $\alpha = 0.25$)	Y (Inv. Freq)	Deeper
S6	FL ($\gamma = 2$, $\alpha = 0.25$)	N	Baseline

Categorical Cross-Entropy (CCE) is used as the baseline loss function, measuring the divergence between the predicted probability distribution and true labels using the formula [31]. CCE is calculated as:

$$L_{CCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log(p_{ik}) \quad (3)$$

where y_{ik} is an indicator function (1 if sample i belongs to class k, 0 otherwise) and p_{ik} is the predicted probability for class k.

Focal Loss (FL) is applied to address class imbalance by down-weighting easy-to-classify examples and increasing focus on hard examples through modulating factors $1 - p_{ik}$ [32]. FL is calculated as:

$$L_{FL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k(1 - p_{ik})^{\gamma} y_{ik}\log(p_{ik}) \quad . (4)$$

where the focusing parameter $\gamma = 2$ and class-balancing weight $\alpha = 0.25$ are chosen based on the standard configuration that has been proven effective in the literature for addressing extreme class imbalance [32]. The comparison between CCE and FL was intended to assess whether FL increases sensitivity to hard classes [17], [20].

In addition to modifying the loss function, inverse frequency-based class weighting is applied to increase the importance of underrepresented classes [33]. The class weights are calculated as:

$$w_k = \frac{N}{K \cdot n_k} \quad (5)$$

where N is the total samples, K is the total classes, and n_k is class $k's$ count. Class weights w_k multiply each class's loss contribution, giving hard classes greater training influence [34]. This loss-weighting combination is evaluated via an ablation study to isolate its effects on macro F1 and mean hard-class F1 [18].

In addition, the configurations of the three classifier-head variants evaluated in Phase 4 can be seen in Figure 5.

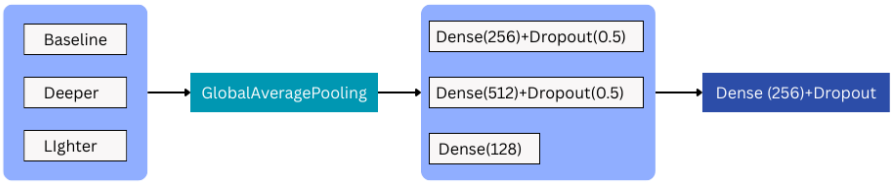

Figure 5. Ablation study classifier head configurations

GlobalAveragePooling2D (GAP) replaces Flatten as illustrated in Figure 5, to reduce parameters and overfitting while preserving spatial information. GAP averages each feature map into a vector, reducing computational complexity and serving as a structural regulariser that forces feature map confidence in the target class. We used a dropout rate of 0.5 in order to reduce overfitting and improve generalisation; and used ReLU activation to avoid vanishing gradients and accelerate training [35].

Experimental Setup and Model Architecture

The lightweight backbone architecture selected was the MobileNetV3-Small (1.1M parameters, 4.18 MB), due to its strong performance in edge and resource-constrained environments [1]. We also compared it with EfficientNetB0 to determine whether these improvements could be applied to larger models.

We initialised both models with ImageNet weights and then fine-tuned them in two stages. We froze the backbone and trained the classifier head only for five epochs with a learning rate of 1e-3 as an initial check, then unfroze the last 20 layers and trained them for 30 epochs using a learning rate of 1e-5 to avoid catastrophic forgetting [36].

We trained with Adam optimiser using a batch size of 32 and an input resolution of 224×224 pixels. To stabilise training and prevent overfitting, we used three standard callbacks: Early Stopping

(five-epoch patience), Model Checkpoint (save best model by validation accuracy), and ReduceLROnPlateau (halve the learning rate after three epochs without improvement).

We used a two-level experimental framework (Table 8) to balance compute and statistical reliability. The main metric was mean F1 across the 12 hard classes; we also report global accuracy, macro F1, and weighted F1. For Tier-2, we additionally report standard deviation (SD), 95% confidence intervals (CIs) (t-distribution, df = 2), and paired t-test p-values. All results were computed on the held-out test set (12% split) to ensure unbiased generalisation estimates.

Table 8. Two-tier experimental design

Tier	Configurations tested	Replications
Tier 1 (Exploratory)	All ablation scenarios	Single run, seed = 1337
Tier 2 (Confirmatory)	Top 3 from Tier-1: (i) best CCE, (ii) best focal loss, (iii) EfficientNet with S0	3 seeds (21, 189, 1337)

RESULT AND DISCUSSION

Impact of IQA and Targeted Augmentation

Applying IQA (LV < 150 or SE < 6.4) to the training images of the 12 hard classes eliminated between 0.3% (C30) and 28.1% (C31) of samples. Targeted offline augmentation of the IQA-passed images increased each hard class to approximately 9,000–12,000 training samples, while the 30 major classes remained at their original counts (approximately 2,000 each). This preserved a realistic, partially imbalanced distribution.

Table 9. Baseline vs. S0 (MobileNetV3-Small) on the NPD test set

Metric	Baseline	S0	Change
Test accuracy	0.9670	0.9808	+0.0138
Macro F1	0.9667	0.9807	+0.0140
Mean F1 (12 hard classes)	0.9260	0.9668	+0.0408

The baseline model was then retrained on this enhanced dataset using the simplest configuration: categorical cross-entropy (CCE), no class weights, and the baseline classifier head (denoted S0). Table 9 compares the baseline and S0 performance. These values were derived from the training logs: seed 1337 for baseline and seed 1337 for S0.

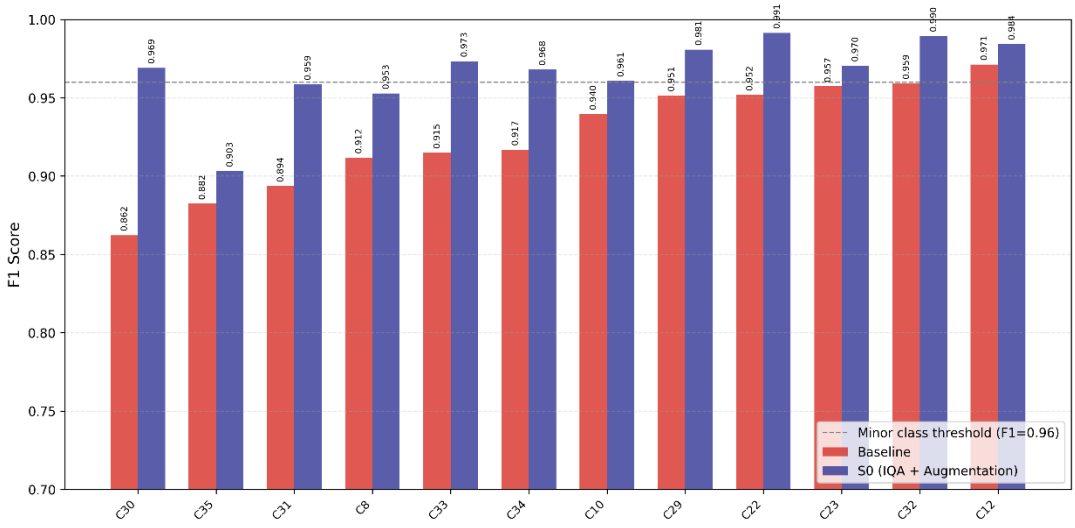

Figure 6. F1 Specific Class Improvement: Baseline vs. S0 (12 hard classes)

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

The average F1 score for the 12 hard classes increased by 4.08 percentage points, while the global accuracy and macro F1 showed a slight improvement. Figure 6 illustrates the increase in F1 scores by class, and Table 10 presents a complete breakdown of the numerical F1 scores for all ablation scenarios. Each of the hard classes showed measurable progress; the largest increases occurred in C30 (+0.1068), C31 (+0.0650), and C33 (+0.0584). Only C35 remained below 0.9100, although its score increased from 0.8824 to 0.9032, and it remains the persistent bottleneck. Overall, our data-centric pipeline substantially improved hard-class robustness without compromising overall performance; while we observed a moderate improvement in overall metrics as well. All values corresponded to Tier 1 (seed = 1337) under the 68:20:12 stratified split.

Table 10. F1 scores for the 12 hard classes across all Tier 1 scenarios (seed = 1337)

Class	Baseline	S0	S1	S2	S3	S4	S5	S6	Delta F1 (S0-Baseline)
C8	0.9118	0.9528	0.9528	0.9325	0.9495	0.9553	0.9206	0.9467	+0.0410
C10	0.9396	0.9607	0.9586	0.9444	0.9582	0.9567	0.9340	0.9571	+0.0211
C12	0.9712	0.9843	0.9876	0.9876	0.9822	0.9930	0.9821	0.9790	+0.0131
C22	0.9519	0.9914	0.9845	0.9707	0.9810	0.9880	0.9654	0.9880	+0.0395
C23	0.9573	0.9704	0.9704	0.9721	0.9627	0.9669	0.9647	0.9612	+0.0131
C29	0.9515	0.9808	0.9865	0.9788	0.9883	0.9865	0.9826	0.9844	+0.0293
C30	0.8623	0.9691	0.9561	0.9386	0.9489	0.9650	0.9343	0.9686	+0.1068
C31	0.8936	0.9586	0.9384	0.9088	0.9345	0.9429	0.9084	0.9567	+0.0650
C32	0.9593	0.9895	0.9947	0.9894	0.9860	0.9877	0.9825	0.9947	+0.0302
C33	0.9149	0.9733	0.9711	0.9651	0.9711	0.9651	0.9595	0.9772	+0.0584
C34	0.9168	0.9680	0.9655	0.9656	0.9617	0.9660	0.9547	0.9734	+0.0512
C35	0.8824	0.9032	0.8786	0.8665	0.8605	0.8728	0.8622	0.8889	+0.0208
Mean	0.9260	0.9668	0.9621	0.9517	0.9570	0.9622	0.9459	0.9647	+0.0408

Ablation Study: Loss Functions, Class Weighting, and Classifier Heads

We conducted a single-seed ablation study (1337) across seven configurations (S0-S6). Table 11 displays the average hard-class F1 scores for each scenario. While incorporating class weights based on CCE (S1) raised the average F1 score from 0.9260 to 0.9621, it still fell short of the standard CCE baseline without weighting (S0). Combining focal loss with weighting (S3) resulted in only a very limited improvement (0.9570). Even without weighting, focal loss alone (S6) reached 0.9647, still below the standard CCE baseline.

Regarding classifier head design, the deeper head (S2) reduced the average F1 score of the hard classes to 0.9517. Meanwhile, the lighter head (S4) yielded a score of 0.9622 but resulted in a trade-off between precision and recall. In contrast, the basic head (S0) delivered the best overall result, with an average hard-class F1 score of 0.9668. These findings indicate that a moderate-capacity head that uses dropout regularisation provides an appropriate balance, which is flexible enough to learn meaningful representations yet has sufficient regularisation to avoid overfitting.

Table 11. Average F1 score for the 12 hard classes from the Tier 1 ablation study (seed = 1337)

Scenario	Test Acc. (%)	Macro F1	Mean F1 (12 hard classes)	Ranking
S0	98.08	0.9807	0.9668	1
S1	98.25	0.9823	0.9621	4
S2	97.80	0.9776	0.9517	6
S3	98.03	0.9800	0.9570	5
S4	98.29	0.9826	0.9622	3
S5	97.59	0.9755	0.9459	7
S6	98.12	0.9810	0.9647	2

Two-Tier Validation and Statistical Significance

To ensure reproducibility and statistical reliability, we repeated the two best configurations (MobileNetV3-Small S0 and S6) as well as EfficientNetB0 utilising the same S0 setup across three random seeds (21, 1337, and 189). Table 12 presents the average and standard deviation of the F1 scores for each hard class.

Paired t-tests showed no significant difference between MobileNetV3-Small S0 and S6 ($p = 0.230$), nor between MobileNetV3-Small S0 and EfficientNetB0 S0 ($p = 0.089$), whereas all three significantly outperformed the baseline ($p = 0.001$ for S0 vs. baseline). These results indicate that the lightweight MobileNetV3-Small with the simple S0 configuration performs on par with both the heavier EfficientNetB0 and the focal loss variant.

Table 12. Tier-2 Validation Results

Configuration	Mean F1	SD	95% CI	vs. S0 p-value
MobilenetV3-Small S0	0.9648	0.0019	[0.960–0.970]	-
MobilenetV3-Small S6	0.9636	0.0010	[0.961–0.966]	0.230
EfficientNetB0 S0	0.9673	0.0008	[0.965–0.969]	0.089

Discussion

Interpretation of Key Findings

Our results highlight four key points. First, interventions targeting data quality are proven to be more effective than architectural adjustments. Simply switching from the default baseline to S0, without changing the loss function, class weights, or classifier head, raised the F1 score of the hard classes by 4.08 percentage points. This finding indicates that targeted IQA filtering, combined with class-specific augmentation, is sufficient to close most of the performance gap on hard classes [8]. Second, class weighting actually has a negative effect after data augmentation. The performance of S1 (with weighting) is 0.0047 lower than that of S0 (no weighting) in terms of the mean hard-class F1 score, because weighting excessively amplified the hard classes, which have already been rebalanced, and distorted the loss landscape. Thirdly, with the standard focal loss configuration ($\gamma=2$, $\alpha=0.25$), we found no advantage over standard CCE after data augmentation (S6: 0.9647 vs. S0: 0.9668). When combined with weighting, this method further reduced performance (S3: 0.9570 vs. S1: 0.9621). Fourth, the baseline classification head proved to be optimal as deeper heads suffered from overfitting (as indicated by a sharp drop in recall), while less complex heads disrupted the balance between precision and recall.

Comparison with Prior Work on the NPD Dataset

To provide context for our findings, we compared them with three recent studies that used the NPD dataset. Priambodo et al. [26], our previous work, evaluated MobileNetV3-Small under standard training and reported a macro-F1 score of approximately 0.97; although it provided per-class F1 score, it only discussed those hard classes in general terms. Imanulloh et al. [22] achieved 97–98% overall

accuracy with a custom-designed CNN, but their confusion matrix indicated some classes with F1 scores below 0.90. Sun et al. [20] implemented EfficientNet with focal loss on a 10-class subset of NPD and reported precision and recall scores for each class ranging from 0.91 to 0.99, with the lowest performance consistently observed on corn and tomato diseases when using the baseline ResNet50 model. Conversely, our S0 configuration achieved an average hard-class F1 score of 0.9668 on our 68:20:12 stratified split. Direct comparisons with prior studies remain limited due to differences in training/test splits and evaluation protocols. Nevertheless, our result demonstrates competitive performance on categories classified as hard classes. To the best of our knowledge, no previous study on NPD has explicitly defined and tracked the average F1 score across a set of hard classes previously identified as systematically difficult to clasify, and then used this as the primary evaluation metric.

Qualitative Examples

To reinforce the quantitative analysis, we also visually examined selected test samples. Figure 7 illustrates a case where S0 produced correct predictions but with low confidence level (< 0.6), reflecting the uncertainty that still exists on challenging images. Figure 8 shows the opposite pattern: high confidence level (> 0.9), but incorrect predictions, which may indicate spurious correlations or annotation ambiguities.

These visual examples reinforce our quantitative findings; although the data-centric pipeline improved overall performance, some challenging samples still resulted in correct predictions with low confidence or errors with high confidence. We regard such cases as suitable candidates for further data inspection or annotation refinement.

Figure 7. Examples of correct predictions with low confidence level (< 0.6)

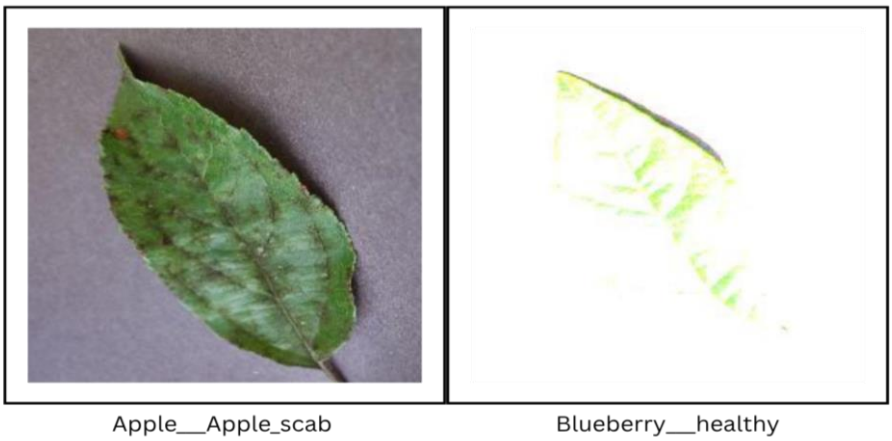

Figure 8. Examples of false predictions with high confidence level (> 0.9)

Why Do C35, C31, and C8 Remain Persistent Bottlenecks?

In all the scenarios we tested, three classes consistently emerged as bottleneck: C35, C31, and C8. Table 13 shows that C35 persistently recorded the lowest F1 scores across all configurations, ranging from 0.860 to 0.903; even with the heavier EfficientNetB0 backbone, its score reached only 0.908. From the confusion matrix, we observed that C35 was most frequently misidentified as C34 and C38. This pattern is consistent with visual observation: target spots in the early stages can resemble small necrotic lesions similar to mite damage, whereas in the advanced stages can be mistaken for healthy spots. The situation is complicated by the fact that C35 exhibits high intra-class variation (lesions can range from tiny dots to large concentric rings) as well as low inter-class separation from C34.

Table 13. The top three persistent bottleneck classes and their primary misclassification patterns (MobileNetV3-Small S0, seed=1337)

Code	S0 F1	Most frequent misclassification	Count	Visual similarity rationale
C35	0.9032	C34	5	Necrotic lesions resemble damage caused by mites
C35	0.9032	C38	10	Early stage lesions resemble healthy patches
C31	0.9586	C9	6	Concentric lesions that visually overlap
C8	0.9528	C10	13	Elongated necrotic lesions with the same morphology

There were also issues with C31 and C8, although not as severe as with C35. C31 was often misclassified as C9, while C8 was confused with C10 in 13 test images. In both cases, the issue appears to arise from overlapping visual features, specifically concentric ring patterns for the tomato blights and elongated necrotic lesions for the corn diseases. This suggests that visual similarity between certain disease classes may set a natural limit on how well lightweight CNNs can distinguish them.

Finally, the persistent difficulty of these three classes suggests that MobileNetV3-Small may not have sufficient representational capacity to capture the fine-grained visual details needed to reliably separate them. Even with data enrichment, the model may lack the capacity to learn the fine-grained visual details needed to reliably separate these classes. Given that the primary bottleneck appears to be data-related, further improvements to data quality should take precedence over architectural modifications such as attention mechanisms.

Risk of Overfitting to Augmented Synthetic Samples

Our targeted augmentation increased the training samples of hard classes five-fold. While the validation-test gap remained very small (<0.2%), indicating low overfitting to the specific augmentation distribution, there remains a risk that the model learns synthetic artefacts rather than true morphological variations. For example, the consistent improvement on C30 across all scenarios might partly reflect that the augmentation aligns well with real-world variations for that disease. However, without testing on an out-of-domain field-captured dataset (e.g., PlantDoc), we cannot guarantee that the gains generalise beyond the NPD test set, which shares the same controlled background as the training images.

Relation to Data-Centric AI Principles

These findings are consistent with the data-centric AI philosophy advanced in the data-centric AI literature, namely the premise that improving data quality and consistency is more important than refining model architecture[37]. The main conclusion of our study, that once data has been cleaned and enriched, using standard CCE with a basic head actually outperforms more sophisticated loss functions, further reinforces the notion that improvements in data quality may offer greater return than increasing algorithmic complexity when data itself is the limiting factor. Moreover, our four phase pipeline

employs a systematic, data-driven procedure specifically designed for plant leaf disease classification. The counterintuitive finding that class weighting becomes detrimental after further enrichment underscores the need to re-evaluate standard imbalance-mitigation techniques after the dataset has been enhanced, a nuance often overlooked in data-centric learning studies.

Threats to Validity

- **Internal validity:** We used fixed IQA thresholds (LV = 150 and SE = 6.4) and a stratified split to reduce procedural bias. However, the relative contribution of IQA filtering and targeted augmentation cannot be fully separated in the current design. Because IQA scores were computed over the whole image, including the dark background around NPD leaves, background effects may also have influenced the quality measures. Future work should include a no-IQA control and background masking to isolate these effects more clearly.
- **External validity:** The findings are limited to the NPD dataset. Whether the proposed pipeline generalises to field images with uncontrolled lighting, backgrounds, and acquisition conditions remains an open question for future work.
- **Construct validity:** Hard classes were defined as classes with F1 < 0.96 or recall < 0.96 based on the baseline performance distribution and the diagnostic risk of false negatives. Although this threshold was empirically motivated, other threshold choices could yield a different set of hard classes and should be examined in future studies.
- **Conclusion validity:** The confirmatory experiment used three random seeds, which limits statistical power. Nevertheless, the rankings remained consistent across seeds, and the standard deviation was below 0.0025. We did not use k-fold cross-validation because of computational constraints; therefore, the reported statistical tests should be interpreted as supporting evidence rather than definitive proof.

CONCLUSION

Our data-centric validation improves robustness for hard classes in lightweight CNN-based plant leaf disease classification. On the NPD dataset with MobileNetV3-Small, the average F1 for the 12 hard classes increased from 0.926 (baseline) to 0.967 (S0), a 4.1 percentage-point gain, while overall test accuracy remained stable at 98.08%. A two-tier, three-seed experiment confirmed robustness (mean hard-class F1 = 0.9652 ± 0.0020).

Several findings differed from what we expected. Inverse class weighting reduced the F1 score for difficult classes from 0.9668 to 0.9621. The focal loss function performed slightly worse than cross-entropy loss (S6: 0.9647 vs. S0: 0.9668). The baseline classifier head outperformed both deeper and lighter architectures. MobileNetV3-Small S0 and EfficientNetB0 were statistically equivalent (p=0.089). C35 has unique characteristics; we observed that this class is the most persistent bottleneck across all scenarios, with an F1 score consistently below 0.91.

For future work, we suggest applying Grad-CAM to diagnose C35 failures, exploring MobileViT and EfficientFormer for fine-grained feature extraction, extending this approach to YOLO-nano and SSD for localisation, and validating the workflow on the PlantDoc dataset to assess its generalisation ability under real-world field conditions.

SUPPLEMENTARY MATERIALS

Additional results, including a complete sensitivity analysis, a confusion matrix, details on classification errors, and an F1 score table by class are available in the supplementary files.

ACKNOWLEDGMENT

All data used in this work were obtained from a publicly available source, and no human or animal subjects were involved. The authors declare that they have no conflicts of interest.

REFERENCES

[1] A. T. Khan, S. M. Jensen, A. R. Khan, and S. Li, "Plant disease detection model for edge computing devices," *Front. Plant Sci.*, vol. 14, p. 1308528, Dec. 2023, doi: 10.3389/fpls.2023.1308528.
[2] J. Yang, G. Xu, M. Yang, and Z. Lin, "Lightweight wavelet-CNN tea leaf disease detection," *PLOS One*, vol. 20, no. 5, p. e0323322, May 2025, doi: 10.1371/journal.pone.0323322.
[3] U. V. Nnamdi and V. Abolghasemi, "Optimised MobileNet for very lightweight and accurate plant leaf disease detection," *Sci. Rep.*, vol. 15, no. 1, p. 43690, Dec. 2025, doi: 10.1038/s41598-025-27393-z.
[4] J. Zhang, X. Yang, X. Fu, B. Wang, and H. Li, "LDL-MobileNetV3S: an enhanced lightweight MobileNetV3-small model for potato leaf disease diagnosis through multi-module fusion," *Front. Plant Sci.*, vol. 16, p. 1656731, Oct. 2025, doi: 10.3389/fpls.2025.1656731.
[5] A. Y. Ashurov *et al.*, "Enhancing plant disease detection through deep learning: a Depthwise CNN with squeeze and excitation integration and residual skip connections," *Front. Plant Sci.*, vol. 15, p. 1505857, Jan. 2025, doi: 10.3389/fpls.2024.1505857.
[6] P. K. Verma, N. Gupta, A. K. Sharma, N. Rakesh, and M. Gulhane, "ViTCon: a hybrid CNN-ViT model for improved plant leaf disease detection," *Cogent Food Agric.*, vol. 11, no. 1, p. 2562165, Dec. 2025, doi: 10.1080/23311932.2025.2562165.
[7] C. Pal, S. Karmakar, I. Mukherjee, and P. P. Chakrabarti, "A lightweight and explainable CNN model for empowering plant disease diagnosis," *Sci. Rep.*, vol. 15, no. 1, p. 30720, Aug. 2025, doi: 10.1038/s41598-025-94083-1.
[8] N. Bhatt, N. Bhatt, P. Prajapati, V. Sorathiya, S. Alshathri, and W. El-Shafai, "A Data-Centric Approach to improve performance of deep learning models," *Sci. Rep.*, vol. 14, no. 1, p. 22329, Sep. 2024, doi: 10.1038/s41598-024-73643-x.
[9] D. Schaudt *et al.*, "Augmentation strategies for an imbalanced learning problem on a novel COVID-19 severity dataset," *Sci. Rep.*, vol. 13, no. 1, p. 18299, Oct. 2023, doi: 10.1038/s41598-023-45532-2.
[10] R. Escobar Díaz Guerrero, L. Carvalho, T. Bocklitz, J. Popp, and J. L. Oliveira, "A Data Augmentation Methodology to Reduce the Class Imbalance in Histopathology Images," *J. Imaging Inform. Med.*, vol. 37, no. 4, pp. 1767–1782, Mar. 2024, doi: 10.1007/s10278-024-01018-9.
[11] D. Dablain, C. Bellinger, B. Krawczyk, and N. Chawla, "Efficient Augmentation for Imbalanced Deep Learning," arXiv:2207.06080, 2022, doi: 10.48550/arXiv.2207.06080.
[12] Y. Yang, C. Liu, H. Wu, and D. Yu, "A quality assessment algorithm for no-reference images based on transfer learning," *PeerJ Comput. Sci.*, vol. 11, p. e2654, Jan. 2025, doi: 10.7717/peerj-cs.2654.
[13] Q.-H. Miao *et al.*, "Deep learning-based no-reference quality assessment of anterior segment ultrasound biomicroscopy panoramic images," *Quant. Imaging Med. Surg.*, vol. 15, no. 11, pp. 11219–11236, Nov. 2025, doi: 10.21037/qims-2025-592.
[14] D. J. Richter and K. Kim, "Assessing the performance of domain-specific models for plant leaf disease classification: a comprehensive benchmark of transfer-learning on open datasets," *Sci. Rep.*, vol. 15, no. 1, p. 18973, May 2025, doi: 10.1038/s41598-025-03235-w.
[15] B. Hu, W. Jiang, J. Zeng, C. Cheng, and L. He, "FOTCA: hybrid transformer-CNN architecture using AFNO for accurate plant leaf disease image recognition," *Front. Plant Sci.*, vol. 14, p. 1231903, Sep. 2023, doi: 10.3389/fpls.2023.1231903.
[16] J. Singh *et al.*, "Batch-balanced focal loss: a hybrid solution to class imbalance in deep learning," *J. Med. Imaging*, vol. 10, no. 05, Jun. 2023, doi: 10.1117/1.JMI.10.5.051809.
[17] M. S. Hossain, J. M. Betts, and A. P. Paplinski, "Dual Focal Loss to address class imbalance in semantic segmentation," *Neurocomputing*, vol. 462, pp. 69–87, Oct. 2021, doi: 10.1016/j.neucom.2021.07.055.
[18] Z. Al Sahili and M. Awad, "The power of transfer learning in agricultural applications: AgriNet," *Front. Plant Sci.*, vol. 13, p. 992700, Dec. 2022, doi: 10.3389/fpls.2022.992700.
[19] A. Ahmad, D. Saraswat, V. Aggarwal, A. Etienne, and B. Hancock, "Performance of deep learning models for classifying and detecting common weeds in corn and soybean production systems," *Comput. Electron. Agric.*, vol. 184, p. 106081, May 2021, doi: 10.1016/j.compag.2021.106081.
[20] X. Sun, G. Li, P. Qu, X. Xie, X. Pan, and W. Zhang, "Research on plant disease identification based on CNN," *Cogn. Robot.*, vol. 2, pp. 155–163, 2022, doi: 10.1016/j.cogr.2022.07.001.
[21] G. A. Sandag, F. M. Tombeng, S. R. Vridvly, J. Waworundeng, and W. Mokodaser, "CornNet: Implementation of Transfer Learning for Disease Classification in Corn Plants Based on Web Application," *CogITo Smart J.*, vol. 11, no. 1, pp. 167–181, Jun. 2025, doi: 10.31154/cogito.v11i1.992.167-181.
[22] S. B. Imanulloh, A. R. Muslikh, and D. R. I. M. Setiadi, "Plant Diseases Classification based Leaves Image using Convolutional Neural Network," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 1–10, Aug. 2023, doi: 10.33633/jcta.v1i1.8877.
[23] N. Ullah *et al.*, "An effective approach for plant leaf diseases classification based on a novel DeepPlantNet deep learning model," *Front. Plant Sci.*, vol. 14, p. 1212747, Oct. 2023, doi: 10.3389/fpls.2023.1212747.

[24] J. S. Aguilar-Ruiz and M. Michalak, "Classification performance assessment for imbalanced multiclass data," *Sci. Rep.*, vol. 14, no. 1, p. 10759, May 2024, doi: 10.1038/s41598-024-61365-z.
[25] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," *Front. Plant Sci.*, vol. 7, p. 1419, Sep. 2016, doi: 10.3389/fpls.2016.01419.
[26] B. Priambodo, A. Fadlil, and S. Sunardi, "Perbandingan CNN untuk Deteksi Penyakit Daun Tanaman New Plant Disease Berbasis Cloud Computing," *J. Inform. Teknol. Dan Sains*, vol. 7, no. 4, 2025, doi: 10.51401/jinteks.v7i4.6782.
[27] J. C. Helps, F. Van Den Bosch, N. Paveley, L. N. Jørgensen, N. Holst, and A. E. Milne, "A framework for evaluating the value of agricultural pest management decision support systems," *Eur. J. Plant Pathol.*, vol. 169, no. 4, pp. 887–902, Aug. 2024, doi: 10.1007/s10658-024-02878-1.
[28] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognit.*, vol. 46, no. 5, pp. 1415–1432, May 2013, doi: 10.1016/j.patcog.2012.11.011.
[29] J. L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Sep. 2000, pp. 314–317 vol.3. doi: 10.1109/ICPR.2000.903548.
[30] L. Silong, Z. Xiaoguang, H. Dongyang, N. Ali, K. Qiankun, and W. Sijia, "A Multi-Feature Framework for Quantifying Information Content of Optical Remote Sensing Imagery," *Remote Sens.*, vol. 14, no. 16, p. 4068, Aug. 2022, doi: 10.3390/rs14164068.
[31] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Comput. Med. Imaging Graph.*, vol. 95, p. 102026, Jan. 2022, doi: 10.1016/j.compmedimag.2021.102026.
[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007. doi: 10.1109/ICCV.2017.324.
[33] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 9260–9269. doi: 10.1109/CVPR.2019.00949.
[34] N. Liu, J. Wang, Y. Zhu, L. Wan, and Q. Li, "Improving imbalance classification via ensemble learning based on two-stage learning," *Front. Comput. Neurosci.*, vol. 17, p. 1296897, Jan. 2024, doi: 10.3389/fncom.2023.1296897.
[35] G. Habib and S. Qureshi, "GAPCNN with HyPar: Global Average Pooling convolutional neural network with novel NNLU activation function and HYBRID parallelism," *Front. Comput. Neurosci.*, vol. 16, p. 1004988, Nov. 2022, doi: 10.3389/fncom.2022.1004988.
[36] S. M. Hassan, A. K. Maji, M. Jasiński, Z. Leonowicz, and E. Jasińska, "Identification of Plant-Leaf Diseases Using CNN and Transfer-Learning Approach," *Electronics*, vol. 10, no. 12, p. 1388, Jun. 2021, doi: 10.3390/electronics10121388.
[37] M. H. Saleem, F. Hammad, M. Taha, S. Palaiahnakote, S. U. Rehman, and M. Saraee, "A novel data-centric AI approach based on sensitivity and correlation analyses: A case study on multi-organ plant disease classification," *Expert Syst. Appl.*, vol. 290, p. 128365, Sep. 2025, doi: 10.1016/j.eswa.2025.128365.