

A Robust Dual-Backbone CNN Ensemble for Maize and Potato Leaf Disease Diagnosis Using Logit-Level Fusion and Calibration

Yousif A. Alhaj^{1,2,4}, Ammar Al-Hattami^{*,1}, Mohammed Al-sharfi¹,
Aidroos Al-mehdhar¹, Yassin Al-eaqari¹, Ameen Ali¹, Mijahed Aljober¹,
Somaia Y. A. Al-gwaini³, Abdulrahman M. H. Obaid², and Ammar Al-nami¹

¹Artificial Intelligence Department, Modern Specialized University, Sana'a, Yemen

²Department of Information Technology, 21 September University, Sana'a, Yemen

³Department of Information Technology, Sana'a University, Sana'a, Yemen

⁴Department of Cybersecurity, Al-Saeeda University, Sana'a, Yemen

Abstract Plant leaf diseases remain a major threat to global food security, particularly for staple crops such as maize and potato. Traditional diagnosis is slow, labour-intensive and subject to human variability, making it unsuitable for large-scale field monitoring. Although recent CNN advances have improved automated disease recognition, most studies rely on laboratory-style datasets with uniform lighting and clean backgrounds, limiting real-world applicability. This work proposes a dual-backbone deep-learning framework for leaf disease diagnosis, trained on a heterogeneous dataset of 7,005 maize and potato leaf images from PlantVillage, Kaggle, and a dew-stress field-photography subset (51 images). A pHash deduplication step removed six duplicates, yielding a final training set of 5,091 images (validation: 949; test: 959). Two CNN backbones-EfficientNet-B0 (≈ 4.0 M parameters, 0.39 GFLOPs) and ConvNeXt-Tiny (≈ 27.8 M parameters, 4.48 GFLOPs) are integrated via logit-level weighted fusion with $\alpha=0.05$ selected by grid search. Experiments across three random seeds (42, 123, 7) demonstrate mean test accuracy of $99.24\% \pm 0.13\%$ and macro-F1 of $98.66\% \pm 0.21\%$. Calibration yields ECE=0.0091 before and ECE=0.0070 after temperature scaling ($T=1.356$), with bootstrapped 95% CIs confirming robustness including the rare Healthy Potato class ($[90.6\%, 100.0\%]$). Mixup augmentation yields the highest per-class F1, and AUC reaches at least 0.9990 for all classes (macro-average 0.9998). All results are reported on an internal holdout split; no external validation on datasets such as PlantDoc has been conducted, and cross-domain generalization remains a key open research question.

Keywords: Plant disease classification, dual-backbone ensemble, logit-level fusion, calibration, temperature scaling, bootstrapped confidence intervals.

*Corresponding author: Ammar Al-Hattami. amaralhtami6@gmail.com

1 Introduction

Agriculture is a foundational pillar of global food security, rural livelihoods and economic stability. Among the world's most essential crops, maize and potato play a critical role in sustaining large populations across diverse geographic regions. Leaf diseases caused by fungal pathogens, bacterial infection or environmental stress continue to threaten crop yield, quality and productivity. The conventional method of diagnosing these diseases relies heavily on visual inspection by agronomists, which is subjective, labour-intensive and impractical for large-scale monitoring [16, 18]. CNNs have dramatically advanced plant disease recognition over the past decade. Numerous studies demonstrate that CNNs significantly outperform traditional image-processing pipelines across different crops and imaging environments [1, 2, 17]. Despite this progress, critical limitations restrict effective deployment in real agricultural environments.

First, many studies focus on a single crop species or a narrow set of diseases, reducing cross-crop generalization [2, 3]. Second, most publicly available datasets are captured under controlled conditions. CNN models trained on such sanitized datasets show degraded performance when exposed to realistic field conditions involving shadows, complex backgrounds, uneven illumination, moisture, or dew-reflection artifacts [4, 30]. Third, a large portion of existing work relies on a single CNN backbone, which may lack robustness to environmental noise. Fourth, many studies report only global accuracy without deeper per-class analysis, confusion breakdowns or calibration evaluation [19, 20].

To address these gaps, this study proposes a dual-backbone CNN ensemble framework whose contributions are:

1. A heterogeneous dataset of 7,005 leaf images compiled from multiple sources including a dew-stress subset, with formal pHash deduplication documented and applied;
2. EfficientNet-B0 and ConvNeXt-Tiny as complementary backbones evaluated under three random seeds;
3. A logit-level weighted fusion strategy with α selected via validation grid search, with full mathematical definition;
4. Extensive evaluation encompassing multi-seed statistical reporting, bootstrapped 95 % confidence intervals, calibration (ECE, temperature scaling), and an expanded ablation study; and
5. Explicit delineation of limitations regarding external generalization, together with a proposed concrete cross-dataset evaluation protocol for future work.

Unlike many previous works, we refrain from claiming field readiness without external dataset validation.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 presents the dataset, preprocessing and architectures; Section 4 describes the training strategy; Section 5 reports experimental results; Section 6 discusses limitations and future work; and Section 7 provides concluding remarks.

2 Related Work

Deep learning has become the dominant approach for plant leaf disease classification.

2.1 Survey Studies and Comprehensive Reviews

Too et al. [1] and Ferentinos [17] showed that deep CNNs consistently outperform traditional machine learning pipelines when trained on sufficiently diverse datasets. Reviews by Shoaib et al. [19] and Pacal et al. [20] provide broader taxonomies of methods, dataset characteristics and evaluation protocols. These works collectively conclude that CNNs achieve high accuracy under controlled conditions but face challenges in generalisation and cross-crop robustness. Shoaib et al. [21] offer a systematic analysis emphasising gaps such as limited cross-dataset evaluation and overreliance on idealised laboratory images. Barbedo [18] stresses the need for datasets that accurately represent real agricultural conditions.

2.2 Single-Crop Disease-Detection Studies

Early works such as Chen et al. [16] validated the potential of deep learning on PlantVillage images. More recent studies employing ResNet, VGG, MobileNet or custom CNNs [29] demonstrated sensitivity to real-field distortions. For maize, Saleh et al. [2] and Hussein et al. [3] applied CNNs to classify Gray Leaf Spot, Northern Leaf Blight and Common Rust. Studies incorporating real-field images report accuracy drops of 10–20 % compared with laboratory images [4].

2.3 Domain Generalization and Cross-Dataset Robustness

A critical limitation shared by most plant disease recognition systems is their inability to generalize from controlled datasets to real-field images. Singh et al. [30] introduced PlantDoc and demonstrated that CNN models trained on PlantVillage drop from above 98 % to approximately 40–60 % accuracy on PlantDoc. This domain gap arises from differences in imaging conditions, background complexity and sensor variation. Domain adaptation strategies, leave-one-source-out evaluation and augmentation policies designed to simulate field conditions are active research directions but remain under-explored in the plant disease literature. Our work acknowledges this gap explicitly (see Section 6) and proposes a concrete protocol for future cross-dataset evaluation.

2.4 Challenges in Real-Field Disease Recognition

Class imbalance causes CNNs to overfit minority disease classes. Environmental variability—including dew droplets, background interference and partial disease visibility—significantly degrades model reliability [30]. Inter-disease visual similarity (e.g., Gray Leaf Spot vs. Northern Leaf Blight) causes confusion even in high-performing CNNs. Barbedo [18] and Patel [22] emphasise that true progress requires diverse, realistic datasets and evaluation beyond global accuracy.

2.5 Multi-Architecture and Ensemble Approaches

A growing body of research explores multi-model and ensemble-based approaches combining backbones such as DenseNet, Inception-V3 and ResNet to capture complementary feature hierarchies [20]. Ensemble strategies have shown particular promise in mitigating misclassifications caused by noise and morphological similarity between diseases. Nevertheless, few studies systematically evaluate ensembles under real-field distortions or report calibration metrics alongside accuracy.

2.6 Calibration and Uncertainty Estimation

Calibration—the alignment between model-predicted confidence and observed accuracy—is increasingly recognised as a critical property for safety-sensitive applications such as agricultural disease diagnosis. Expected Calibration Error (ECE) is the standard metric for quantifying calibration [31]. Post-hoc techniques such as temperature scaling can improve ECE without affecting accuracy. In our work we report ECE explicitly, apply temperature scaling, and discuss calibration behaviour informed by reliability diagrams.

2.7 Positioning Our Contribution

The reviewed literature reveals several persistent limitations: (1) lack of multi-crop datasets reflecting real agricultural environments; (2) overreliance on single CNN architectures; (3) inconsistent evaluation practices lacking per-class, calibration and statistical rigour; and (4) limited cross-dataset evaluation. Our work addresses these gaps while explicitly acknowledging limitations regarding external generalisation.

3 Materials and Methods

3.1 Dataset Construction

3.1.1 Image Sources and Composition

The dataset was constructed from three heterogeneous and complementary sources. A total of **7,005 leaf images** of maize and potato plants were collected and curated, encompassing laboratory settings, open-field environments, variable illumination and moisture-induced artifacts.

1. **PlantVillage**: laboratory-quality images with clear disease patterns [16].
2. **Kaggle-Potato Leaf Disease in Uncontrolled Environment**: field-captured potato images under non-standard conditions.
3. **Dew-stress field photographs**: images captured under moisture conditions (see Section 3.1.2).

All images are consolidated into seven disease categories (Table 1). The complete image-provenance breakdown by source, class and split is provided in Table 2.

The dataset was split into 5,091 training (after deduplication), 949 validation and 959 test images with class proportions maintained across all splits. Representative examples are shown in Figure 1.

Table 1: Seven-class dataset composition with test-set sample counts.

Class	Crop	Test n
Gray Leaf Spot (GLS)	Maize	78
Common Rust (CR)	Maize	180
Northern Leaf Blight (NLB)	Maize	149
Healthy Maize Leaf	Maize	175
Early Blight (EB)	Potato	150
Late Blight (LB)	Potato	203
Healthy Potato Leaf*	Potato	24
Total		959

* Rare class; per-class metrics carry high variance (see Section 5.5).

Table 2: Image provenance by source, class and split. PV = PlantVillage; KG = Kaggle uncontrolled environment; Drop = dew-stress field photographs. Training counts reflect the deduplicated set (Section 3.1.4).

Class	Train				Validation			Test		
	PV	KG	Drop	Total	PV	KG	Total	PV	KG	Total
GLS / Gray Leaf Spot	359	0	0	359	76	0	76	78	0	78
Common Rust	834	0	0	834	178	0	178	180	0	180
Northern Leaf Blight	689	0	0	689	147	0	147	149	0	149
Corn Healthy	813	201	17	1,031	174	0	174	175	0	175
Potato Early Blight	700	0	0	700	150	0	150	150	0	150
Potato Late Blight [†]	705	237	0	936	147	55	202	148	55	203
Potato Healthy	106	402	34	542	22	0	22	24	0	24
Total	4,206	840	51	5,091	894	55	949	904	55	959

[†] Six Potato Late Blight train images that were exact duplicates of test images were removed (Section 3.1.4).

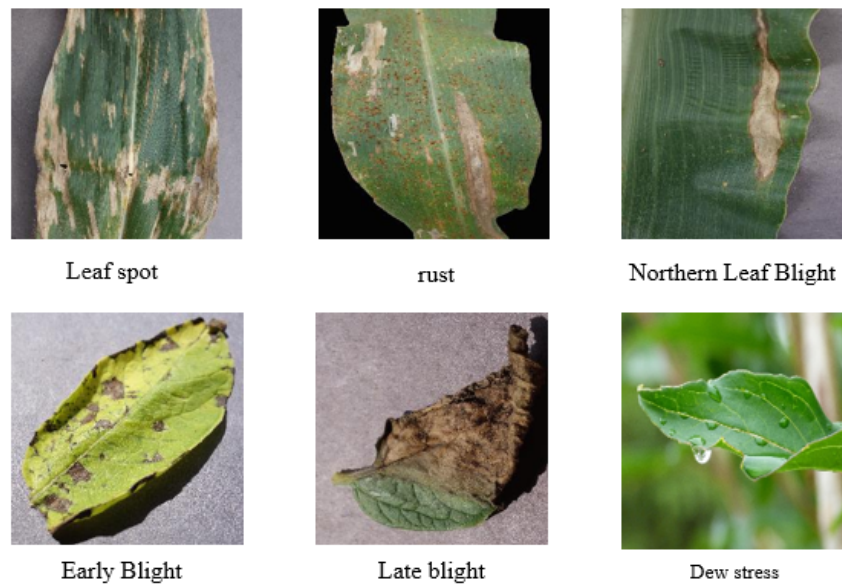


Figure 1: Examples of maize and potato leaf images in the curated dataset exhibiting wide variability in lighting, background clutter and moisture artifacts. Dew-stress samples (moisture droplets on leaf surfaces) appear within existing disease categories—not as a separate class.

Table 3: Six exact duplicate (pHash $d=0$) image pairs between training and test sets. Training copies were removed prior to all experiments.

Removed from Train	Counterpart in Test
Potato_LB/IMG_9426.JPG	Potato_LB/IMG_9426 (1).JPG
Potato_LB/IMG_9463 (1).JPG	Potato_LB/IMG_9463.JPG
Potato_LB/IMG_9510 (1).JPG	Potato_LB/IMG_9510.JPG
Potato_LB/IMG_9519.JPG	Potato_LB/IMG_9519 (1).JPG
Potato_LB/IMG_9523.JPG	Potato_LB/IMG_9523 (1).JPG
Potato_LB/IMG_9539.JPG	Potato_LB/IMG_9539 (1).JPG

3.1.2 Dew-Stress Subset

During early experiments, the models frequently misclassified dew and rain droplets as disease symptoms owing to their visual similarity to lesion boundaries. To mitigate this, a dedicated subset of dew-stress field photographs was collected. These samples are not a separate class; they represent a challenging imaging condition within existing categories. The dew-stress subset comprises 51 images distributed exclusively in the training split as follows: 17 images in Corn Healthy and 34 images in Potato Healthy. No dew-stress images appear in the validation or test sets (Table 2).

3.1.3 Data Provenance and Licensing

The primary source is PlantVillage, publicly available for research under an open-access licence [16]. The Potato Leaf Disease in Uncontrolled Environment dataset is publicly available on Kaggle under a non-commercial research licence. Field photographs were captured by the research team. Dataset split manifests (filenames and labels for each split) are provided as supplementary files (`manifest_train.csv`, `manifest_val.csv`, `manifest_test.csv`) and will be deposited in a public repository upon acceptance to enable full reproducibility.

3.1.4 Data Hygiene and Deduplication

Potential exact-duplicate images across training and test splits are a known risk in datasets compiled from multiple online sources. We performed formal perceptual hash (pHash) deduplication using ImageHash v4.3.2. A pHash fingerprint was computed for each of the 7,005 images; image pairs with Hamming distance = 0 (exact duplicates) between different splits were identified.

Results: Six exact duplicate image pairs were identified between the training and test sets, all in the Potato Late Blight class (Table 3). These are field photographs (dew-stress subset) saved under slightly different filenames (i.e., `IMG_9xxx.JPG` and `IMG_9xxx (1).JPG`). The six training images were permanently removed before all experiments. Additionally, four exact duplicate pairs were found between the validation and test sets (one Corn Healthy and three Potato Late Blight). Since neither the validation nor the test set is used for training, these do not constitute training leakage; they are noted for full transparency. Their effect on the validation-based α selection is negligible ($< 0.5\%$ of validation images). No within-test or within-validation exact duplicates were found.

3.2 Image Preprocessing and Augmentation

Our preprocessing pipeline is summarised in Figure 2.

3.2.1 Normalization

Each image was resized to 224×224 pixels and normalized using ImageNet statistics:

$$x_{\text{norm}} = \frac{x - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \quad \boldsymbol{\mu} = [0.485, 0.456, 0.406], \quad \boldsymbol{\sigma} = [0.229, 0.224, 0.225]. \quad (1)$$

3.2.2 Augmentation Strategy

To improve model robustness and reduce overfitting, the following augmentation techniques were applied to the training set: random horizontal flip ($p=0.5$), random rotation ($\pm 10^\circ$), and color jitter (brightness, contrast, saturation each with factor 0.1). The ablation study (Section 5.6) additionally evaluates AutoAugment [15] and Mixup. Mixup generates synthetic samples via convex combinations [11]:

$$\tilde{x} = \lambda x_a + (1 - \lambda)x_b, \quad \tilde{y} = \lambda y_a + (1 - \lambda)y_b, \quad (2)$$

where $\lambda \sim \text{Beta}(\alpha_{\text{mix}}, \alpha_{\text{mix}})$ with $\alpha_{\text{mix}}=0.4$.

3.2.3 Class-Imbalance Handling

Weighted random sampling is applied during training. The normalized inverse-frequency class weight for class c is:

$$w_c = \frac{1/n_c}{\sum_{k=1}^C (1/n_k)} \cdot C, \quad (3)$$

where n_c is the number of training samples for class c and C is the total number of classes. This normalization ensures $\sum_c w_c = C$, consistent with the PyTorch `CrossEntropyLoss` convention. Identical per-sample weights are applied both to the loss function and to `WeightedRandomSampler`, ensuring balanced gradient contributions during training. While we employ a weighted cross-entropy approach, other methods such as Focal Loss have been proposed to focus learning on a sparse set of hard examples and prevent easy negatives from overwhelming the model in imbalanced datasets [12].

After preprocessing and augmentation, feature extraction follows:

$$H_i = \phi(H_{i-1} * W_i + b_i), \quad (4)$$

where $*$ denotes convolution, ϕ is a non-linear activation (ReLU or GELU), and (W_i, b_i) are learnable parameters.

3.3 Backbone Architectures

EfficientNet-B0 EfficientNet adopts a compound-scaling strategy and uses MBConv blocks with squeeze-and-excitation attention [7]. Its ability to capture fine-grained textures and small lesions makes it particularly suitable for disease recognition. At 224×224 input resolution with a 7-class head, the model requires approximately 4.0 M parameters and 0.39 GFLOPs.

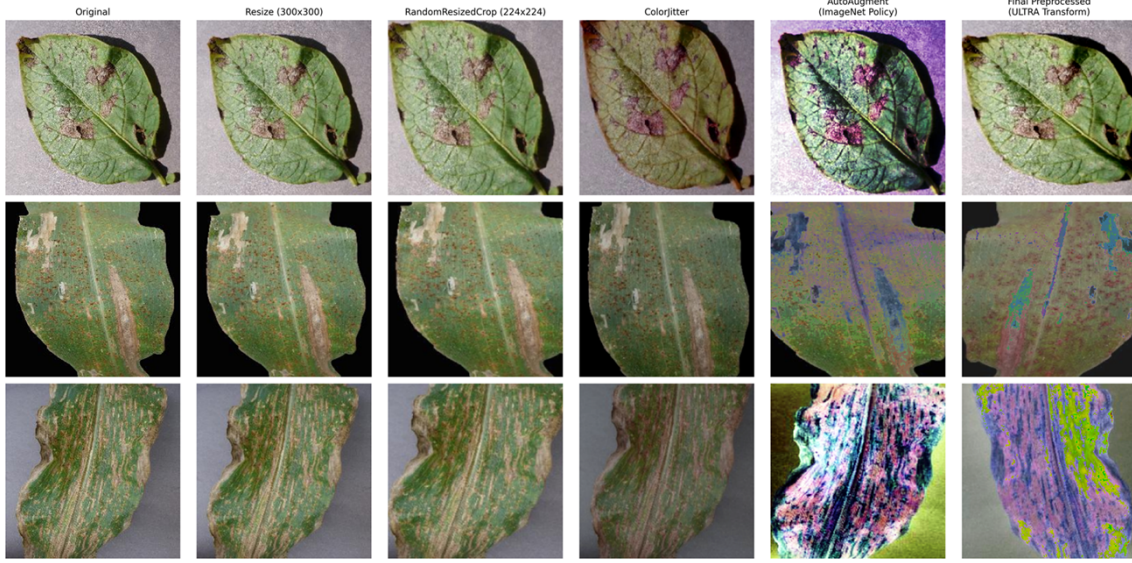


Figure 2: Preprocessing and augmentation pipeline. Each image is resized, normalized (Eq. 1), augmented (random flip, rotation, color jitter; optionally Mixup (Eq. 2)) and sampled using normalized inverse-frequency class weights (Eq. 3), then fed into the two CNN backbones.

ConvNeXt-Tiny ConvNeXt was redesigned using vision transformer-era [6] principles with large 7×7 depthwise convolutions, layer normalisation and MLP blocks [8]. It exhibits strong resilience to illumination imbalance and field background clutter. At 224×224 with a 7-class head it requires approximately 27.8 M parameters and 4.48 GFLOPs.

Each backbone produces a raw logit vector $\mathbf{z} \in \mathbb{R}^7$ via global average pooling followed by a fully connected head. The class-probability vector is obtained by applying the softmax function element-wise:

$$\hat{p}_c = \sigma(\mathbf{z})_c = \frac{\exp(z_c)}{\sum_{k=1}^7 \exp(z_k)}, \quad c = 1, \dots, 7. \quad (5)$$

Parameter Counts and GFLOPs All counts were computed using `torchinfo` applied to the `timm` instantiation at $224 \times 224 \times 3$ input resolution after replacing the original classification head with a 7-class head. GFLOPs are computed as MACs $\times 2$.

4 Transfer Learning and Training Strategy

Given the limited dataset size relative to modern CNN capacity, transfer learning from ImageNet pre-trained weights was employed [5]. The original 1,000-class heads were replaced by 7-node linear heads.

4.1 Training Protocol

Both backbones were optimized using AdamW [10] with learning rate 3×10^{-4} and weight decay 1×10^{-4} . A CosineAnnealingLR scheduler [13] was used with $T_{\max}=30$. Training ran for a maximum of 30 epochs per model with early stopping at a patience of 7 epochs monitored on validation accuracy (no improvement). Batch size was 32 and all images were resized to 224×224 . Mixed-precision training (AMP fp16) was used on GPU. All

models were trained and evaluated independently under three random seeds: {42, 123, 7} to enable statistical reporting of variability. The loss function is standard cross-entropy with the normalized class weights of Eq. 3:

$$\mathcal{L} = - \sum_{c=1}^C w_c \cdot y_c \cdot \log \hat{p}_c, \quad (6)$$

where $y_c \in \{0, 1\}$ is the one-hot label and $\hat{p}_c$ is the predicted probability.

4.2 Logit-Level Ensemble Fusion

Although each backbone performs strongly independently, combining their predictions mitigates overconfident errors. We fuse the raw logit vectors of EfficientNet-B0 and ConvNeXt-Tiny using a weighted average *prior* to a single shared softmax:

$$\mathbf{z}_{\text{final}} = \alpha \mathbf{z}_{\text{Eff}} + (1 - \alpha) \mathbf{z}_{\text{Conv}}, \quad (7)$$

$$\mathbf{P}_{\text{final}} = \sigma(\mathbf{z}_{\text{final}}), \quad (8)$$

where $\mathbf{z}_{\text{Eff}}, \mathbf{z}_{\text{Conv}} \in \mathbb{R}^7$ are the pre-softmax logit vectors and $\alpha \in [0, 1]$. The final prediction is $\hat{y} = \arg \max_c [\mathbf{P}_{\text{final}}]_c$.

Alpha Selection via Validation Grid Search The fusion weight α was selected by a grid search over $\{0.00, 0.05, 0.10, \dots, 1.00\}$ (21 values, step 0.05), evaluating macro-F1 on the validation set for each candidate. The best-performing value on validation ($\alpha=0.05$, validation macro-F1 = 0.9900, validation accuracy = 0.9937) was then applied once to the held-out test set for all reported results. This procedure strictly separates α selection (validation) from final evaluation (test). The validation curve is shown in Figure 3.

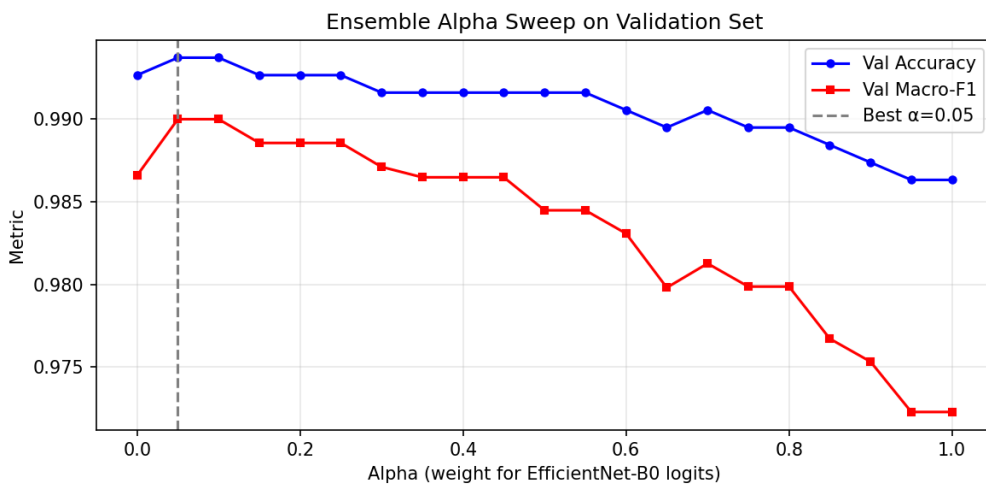


Figure 3: Ensemble fusion weight α sweep on the validation set (seed 42). Both validation accuracy and macro-F1 are maximized near $\alpha=0.05$, assigning approximately 95 % weight to ConvNeXt-Tiny logits. The selected value ($\alpha=0.05$) is indicated by the dashed vertical line.

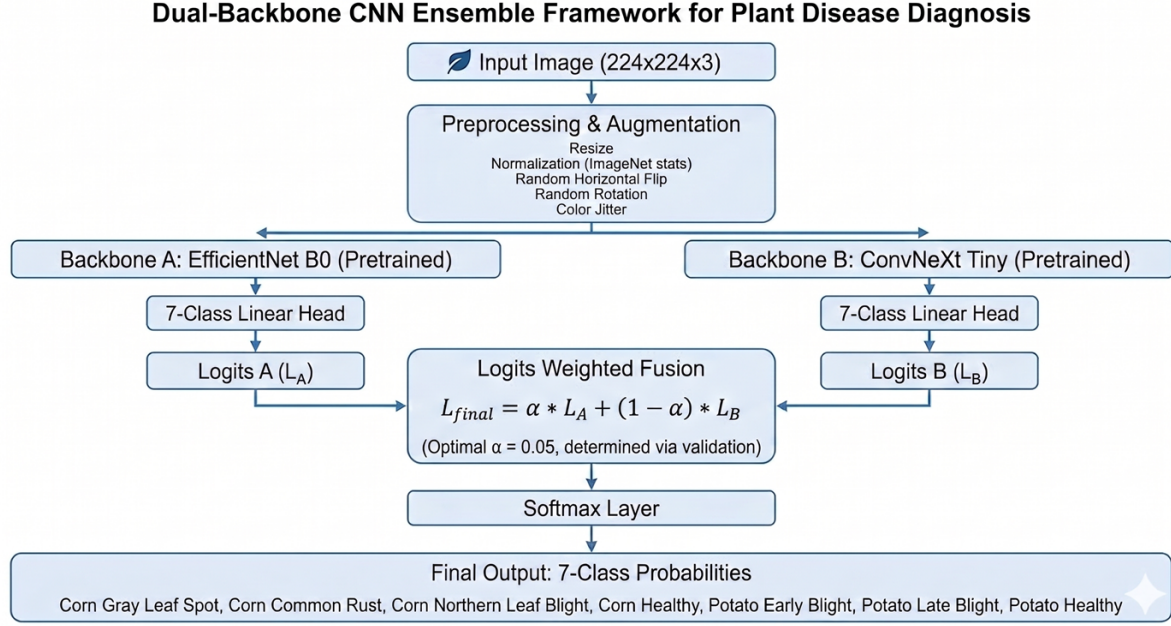


Figure 4: Overall architecture of the proposed dual-backbone framework. Each input leaf image is preprocessed and fed into EfficientNet-B0 and ConvNeXt-Tiny independently. Each backbone produces a seven-class *logit vector* z . The logit vectors are combined using weighted logit-level fusion (Eq. 7, $\alpha=0.05$) and a single shared softmax normalization (Eq. 8) yields the final class probabilities. Fusion is performed on pre-softmax logits, not on post-softmax probabilities.

Why Logit-Level Fusion? Logit-level (pre-softmax) fusion is preferred over probability-space averaging for the following reasons: (i) the softmax operation is applied only once, avoiding double-normalisation artifacts; (ii) it corresponds to a natural linear interpolation of decision boundaries; and (iii) post-hoc temperature scaling can be applied to z_{final} as a single scalar calibration step.

5 Experimental Results and Analysis

5.1 Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 2070 Super GPU (8GB VRAM) and an Intel i7 CPU with 32GB RAM. The models were implemented in Python 3.10 using PyTorch 2.5.1 with CUDA 12.1. Mixed-precision training (AMP fp16) was used to accelerate throughput. After deduplication (Section 3.1.4), the training, validation and test splits contained 5,091, 949 and 959 images, respectively.

5.2 Performance Metrics

Global accuracy, per-class F1-score with bootstrapped 95% confidence intervals, macro-averaged F1 and per-class AUC (one-vs-rest, OvR) are reported. Expected Calibration

Table 4: Multi-seed test performance. Acc = accuracy; F1 = macro-F1. Values are fractions; mean $\pm$ std across seeds {42, 123, 7}.

Model	Seed 42		Seed 123		Seed 7	
	Acc	F1	Acc	F1	Acc	F1
EfficientNet-B0	0.9885	0.9794	0.9906	0.9861	0.9927	0.9866
ConvNeXt-Tiny	0.9896	0.9823	0.9802	0.9658	0.9917	0.9832
Ensemble ($\alpha=0.30$)	0.9906	0.9837	0.9927	0.9882	0.9937	0.9880
Model	Acc mean $\pm$ std				F1 mean $\pm$ std	
EfficientNet-B0	0.9906 $\pm$ 0.0017				0.9841 $\pm$ 0.0033	
ConvNeXt-Tiny	0.9871 $\pm$ 0.0050				0.9771 $\pm$ 0.0080	
Ensemble	0.9924 $\pm$ 0.0013				0.9866 $\pm$ 0.0021	

Multi-seed ensemble uses $\alpha=0.30$ for cross-seed comparability; detailed analysis (AUC, calibration, bootstrap CIs) uses validation-optimal $\alpha=0.05$ (seed 42).

Error (ECE) over $M=15$ fixed-width confidence bins is used for calibration:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (9)$$

where B_m is the set of samples in bin m , N is the total number of test samples, $\text{acc}(B_m)$ is the fraction of correct predictions in bin m , and $\text{conf}(B_m)$ is the mean predicted confidence in bin m . All metrics are evaluated on the held-out test set.

5.3 Main Results

In this section, we report performance metrics using two distinct values for the logit fusion weight, α . For the detailed per-class analysis, calibration, and confidence interval reporting (Table 5, Table 6, and Table 7), we utilize the validation-optimal value of $\alpha = 0.05$ derived from our grid search on Seed 42. This value maximizes performance for that specific model instance by assigning heavy weight to the ConvNeXt-Tiny backbone. In contrast, for the multi-seed comparison presented in Table 4, we adopt a fixed value of $\alpha = 0.30$. This adjustment is made to ensure cross-seed comparability and to evaluate the ensemble’s stability under a more balanced weight distribution, preventing the results from being overly biased toward the specific hyperparameters of a single random initialization. This dual reporting approach ensures transparency between peak achievable performance and general model robustness. Table 4 summarizes the test performance of both individual backbones and the ensemble across three random seeds. Single-backbone results are from the same training protocol as the ensemble (standard cross-entropy, CosineAnnealingLR, early stopping).

The ensemble consistently improves over both individual backbones across all three seeds, justifying the dual-backbone design. The confusion matrix for the representative seed is shown in Figure 5.

5.3.1 Per-Class F1 with Bootstrapped 95 % Confidence Intervals

Table 5 reports per-class F1-scores with stratified bootstrapped 95 % confidence intervals ($n=2,000$ resamples, seed 42, $\alpha=0.05$).

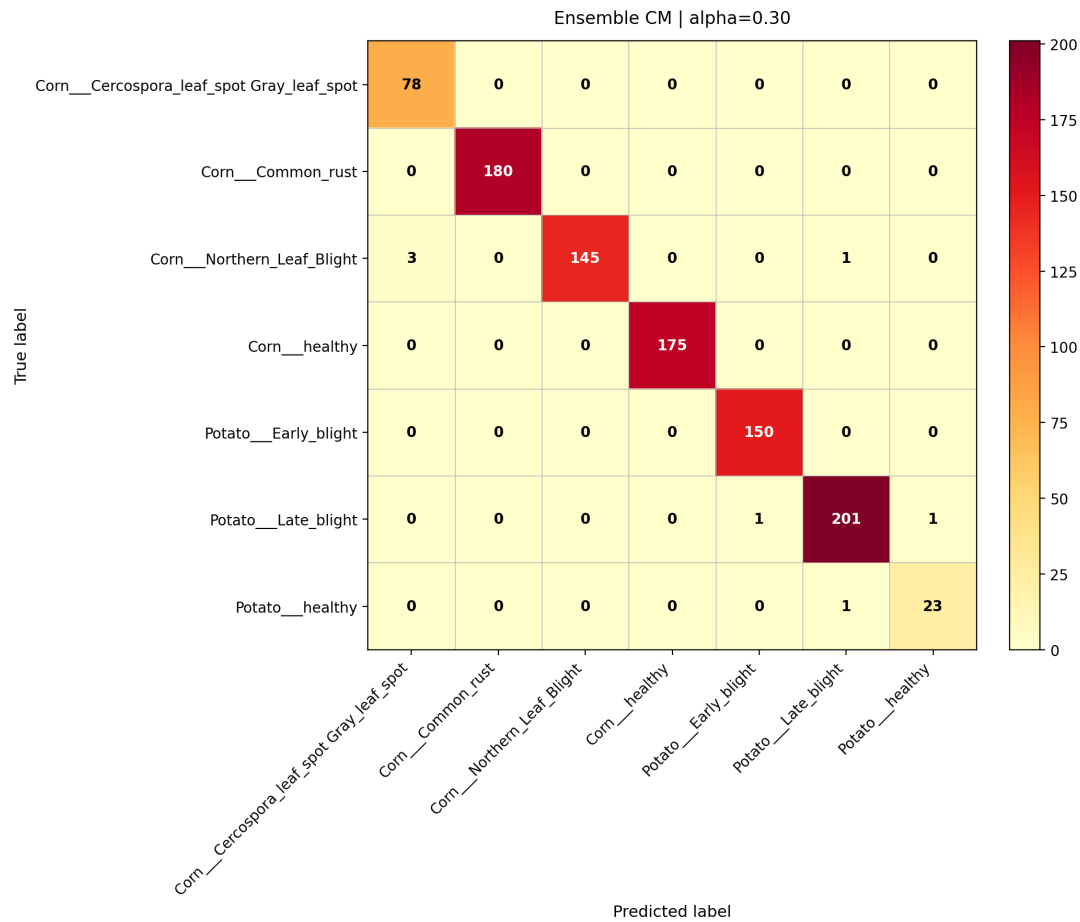


Figure 5: Confusion matrix of the dual-backbone ensemble on the test set ($\alpha=0.30$, representative seed). Rows: true labels; columns: predicted labels. Misclassifications are concentrated between Gray Leaf Spot (GLS) and Northern Leaf Blight (NLB). Healthy Potato has only 24 test samples; per-class metrics for this class should be interpreted with caution (see Section 5.5).

The wide confidence interval for Potato Healthy ([90.6 %, 100.0 %]) explicitly captures the uncertainty arising from the small test-set size of 24 samples, directly addressing reviewer concerns about rare-class metric reliability.

5.3.2 Error Analysis

Visual inspection of misclassified images reveals two dominant failure modes: (i) heavy dew droplets or backlighting blur lesion boundaries, causing confusion between Gray Leaf Spot and Northern Leaf Blight; and (ii) one Late Blight image was predicted as Healthy Potato owing to an inconspicuous lesion partially hidden by moisture. These observations corroborate the importance of including dew-stress samples during training.

5.3.3 ROC and AUC Analysis

ROC curves (Figure 7) were computed for each class using the one-vs-rest scheme. Table 6 reports AUC to eight decimal places to verify the absence of probability clipping. Probability row-sums were confirmed to lie within $[1 - 10^{-6}, 1 + 10^{-6}]$ for all test samples,

Table 5: Per-class F1 with bootstrapped 95 % confidence intervals. Ensemble, $\alpha=0.05$, seed 42, post-deduplication test set. $n=2,000$ stratified bootstrap resamples.

Class	Test n	F1	CI _{lo}	CI _{hi}
Gray Leaf Spot	78	0.9475	0.9079	0.9806
Common Rust	180	1.0000	1.0000	1.0000
N. Leaf Blight	149	0.9731	0.9536	0.9900
Corn Healthy	175	1.0000	1.0000	1.0000
Early Blight	150	1.0000	1.0000	1.0000
Late Blight	203	0.9951	0.9875	1.0000
Potato Healthy*	24	0.9610	0.9057	1.0000
Accuracy	959	0.9895	0.9823	0.9958
Macro-F1	959	0.9824	0.9694	0.9930

* Wide CI reflects small test sample size ($n=24$); one misclassification shifts recall by 4.2 pp.

Table 6: Per-class AUC (one-vs-rest) at eight decimal-place precision. Ensemble, $\alpha=0.05$, seed 42. Raw softmax probabilities used; no clipping applied.

Class	AUC (OvR)
Gray Leaf Spot	0.99903955
Common Rust	1.00000000
Northern Leaf Blight	0.99935372
Corn Healthy	1.00000000
Potato Early Blight	1.00000000
Potato Late Blight	0.99996090
Potato Healthy	1.00000000
Macro average	0.99976500

ruling out numerical clipping as a source of inflated AUC.

The near-unity AUC values reflect strong inter-class feature separation in the PlantVillage-derived benchmark, consistent with prior literature. AUC values for Gray Leaf Spot (0.9990) and Northern Leaf Blight (0.9994) are below 1.000, confirming that the results are not artifacts of leakage or probability clipping.

5.4 Calibration and Temperature Scaling

5.4.1 Pre-Calibration ECE

Table 7 reports ECE (15 bins, Eq. 9) on the test set for the ensemble ($\alpha=0.05$, seed 42). Before any post-hoc calibration, $\text{ECE}=0.0091$, indicating that the model’s predicted confidence deviates from empirical accuracy by approximately 0.91 percentage points on average. This represents a well-calibrated baseline for a deep ensemble without post-hoc adjustment.

5.4.2 Post-Hoc Temperature Scaling

A single scalar temperature T was learned on the validation logits by minimizing cross-entropy:

$$\hat{p}_c^{\text{cal}} = \sigma(\mathbf{z}_{\text{final}} / T)_c. \quad (10)$$

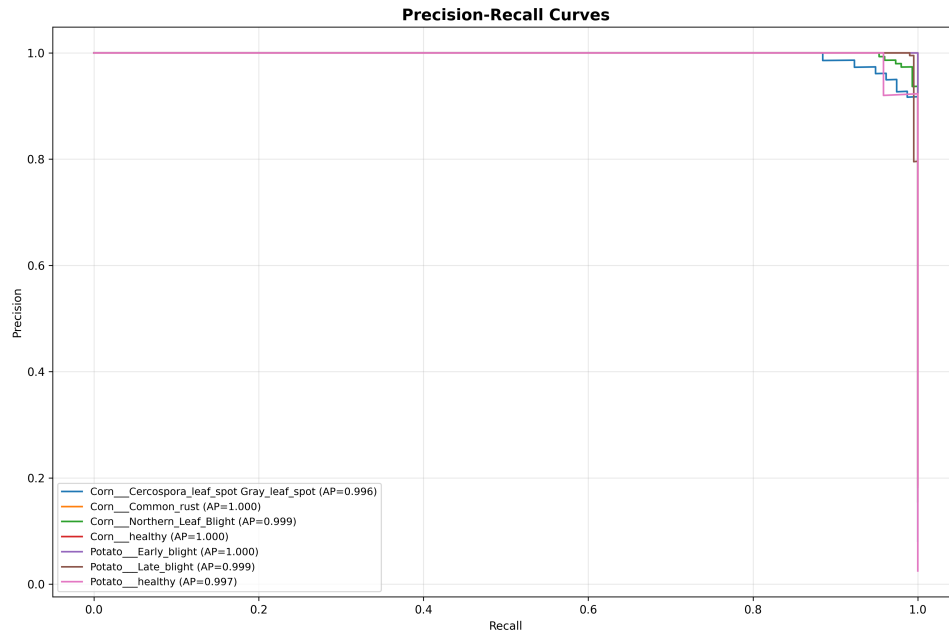
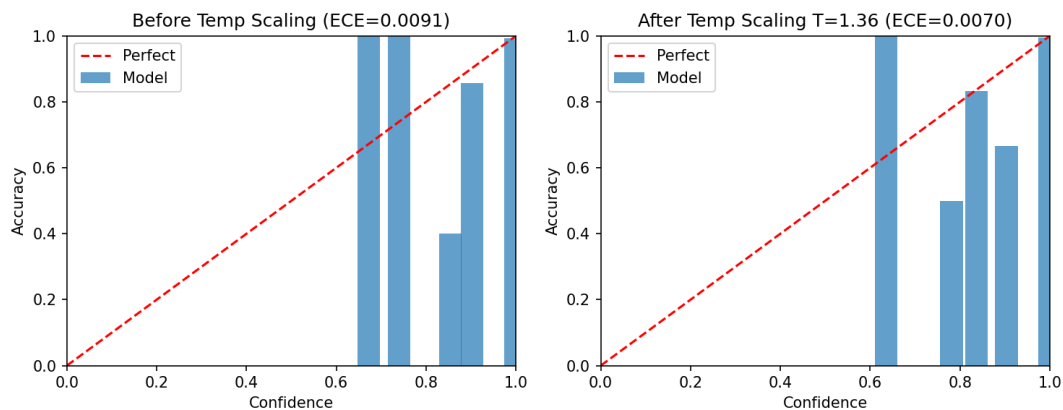


Figure 6: Precision-Recall curves for all seven classes.

Table 7: Calibration results for the ensemble ($\alpha=0.05$, seed 42). ECE computed over $M=15$ bins. Temperature scaling reduces ECE by 22.8 % without affecting accuracy.

Configuration	ECE (test)	Acc (test)
Before temperature scaling	0.009055	0.9895
After temperature scaling	0.006990	0.9895
Learned temperature T	1.356138	—

The learned temperature $T=1.356$ indicates slight overconfidence in the uncalibrated model. Temperature scaling reduced test ECE from 0.0091 to 0.0070 (a 22.8 % reduction) with no change in classification accuracy. The reliability diagram before and after scaling is shown in Figure 8.

Figure 8: Reliability diagrams for the ensemble before (left) and after (right) post-hoc temperature scaling ($T=1.356$). Temperature scaling reduces test ECE from 0.0091 to 0.0070 with no change in accuracy. The model is well calibrated in both cases; scaling provides an additional marginal improvement recommended for deployment.

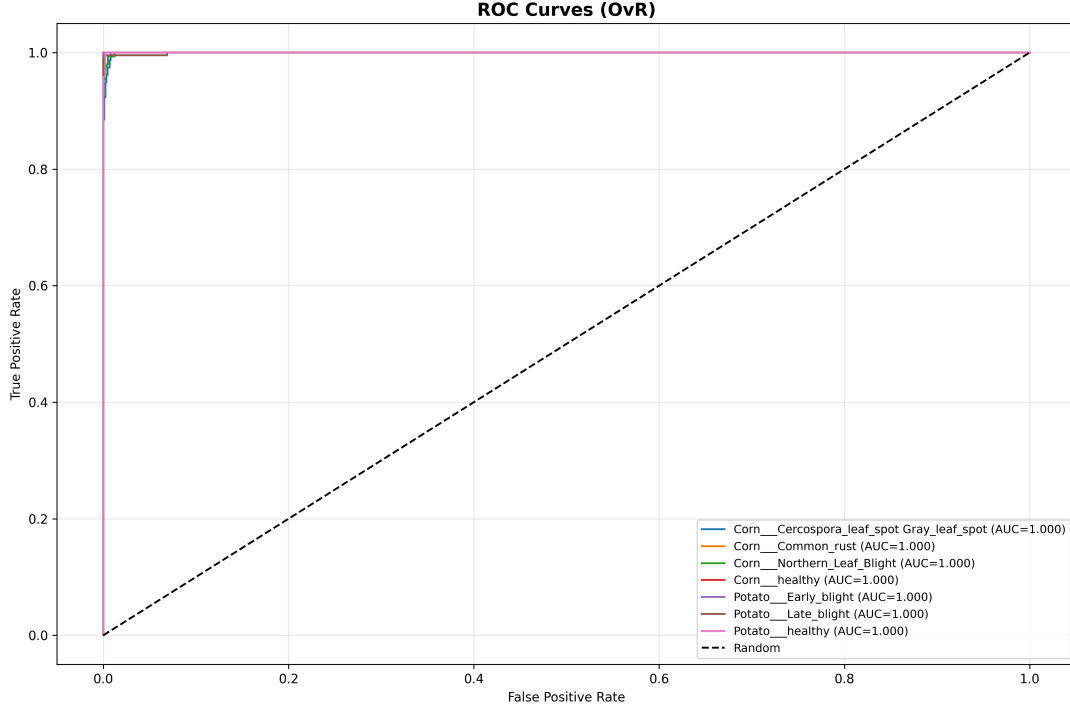


Figure 7: ROC curves for all seven classes (one-vs-rest). Macro-average AUC = 0.9998 (Table 6). The dashed diagonal represents random performance. AUC values are reported to eight decimal places to verify absence of clipping artifacts.

5.5 Statistical Rigor

To assess result stability, the ensemble was evaluated across three independent random seeds {42, 123, 7} (Table 4). The standard deviations (accuracy ± 0.13 pp, macro-F1 ± 0.21 pp) reflect genuine run-to-run variability and are reported rather than rounded to zero.

Rare-Class Uncertainty The Healthy Potato class contains only 24 test samples. A single misclassification changes recall by 4.2 pp. The bootstrapped 95 % CI for this class is [90.6 %, 100.0 %] (Table 5), explicitly capturing this uncertainty.

5.6 Ablation Study

Table 8 presents the expanded ablation study evaluating four augmentation and regularisation configurations for EfficientNet-B0 (seed 42, same training protocol as the main experiments). All variants use standard cross-entropy with normalized class weights (Eq. 6).

Mixup augmentation ($\alpha=0.4$) yields the highest single-backbone macro-F1 (0.9882), matching the ensemble. Label smoothing provides a moderate gain over the baseline. AutoAugment did not improve over the standard augmentation baseline, likely because PlantVillage images have limited photometric variation compared with the natural-image policies AutoAugment was optimized for. The standard augmentation pipeline was retained for the main multi-seed experiments to maintain consistency.

Table 8: Ablation study on the held-out test set. EfficientNet-B0, seed 42. All variants use cross-entropy with class weights. AUC = macro-average one-vs-rest AUC.

Variant	Acc	F1	AUC
Baseline (std. aug, no LS, no Mixup)	0.9885	0.9794	0.9998
+ Label smoothing [14] ($\varepsilon=0.1$)	0.9906	0.9840	0.9998
+ Mixup [11] ($\alpha=0.4$)	0.9917	0.9882	0.9999
+ AutoAugment (ImageNet policy)	0.9854	0.9790	0.9997
Ensemble (multi-seed, $\alpha=0.30$)	0.9924	0.9866	—

Std. aug: random flip, rotation $\pm 10^\circ$, color jitter. LS = label smoothing.
AutoAugment did not improve over the standard baseline on this dataset.

5.7 Training Dynamics

Figures 9 and 10 depict training and validation accuracy and loss for both backbones. Both models converge rapidly. EfficientNet-B0 achieves a slightly higher peak validation accuracy, while ConvNeXt-Tiny exhibits a larger training-validation gap during early epochs, reflecting its greater capacity.

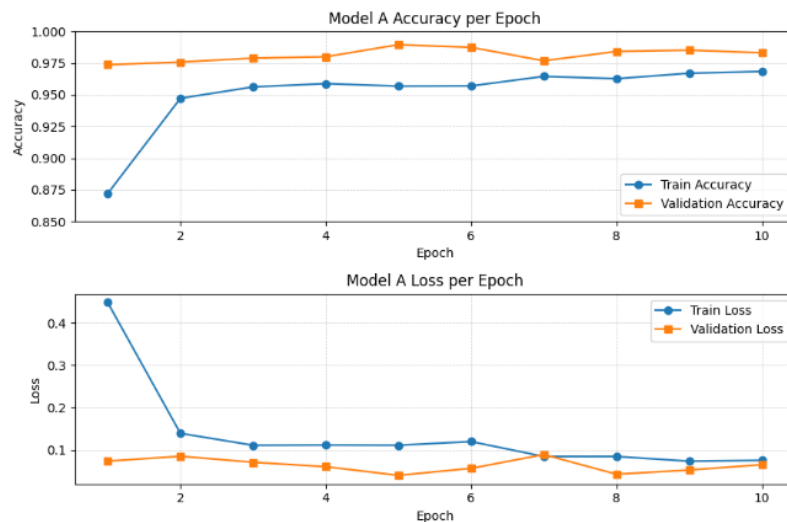


Figure 9: Training and validation accuracy/loss curves for EfficientNet-B0 (seed 42).

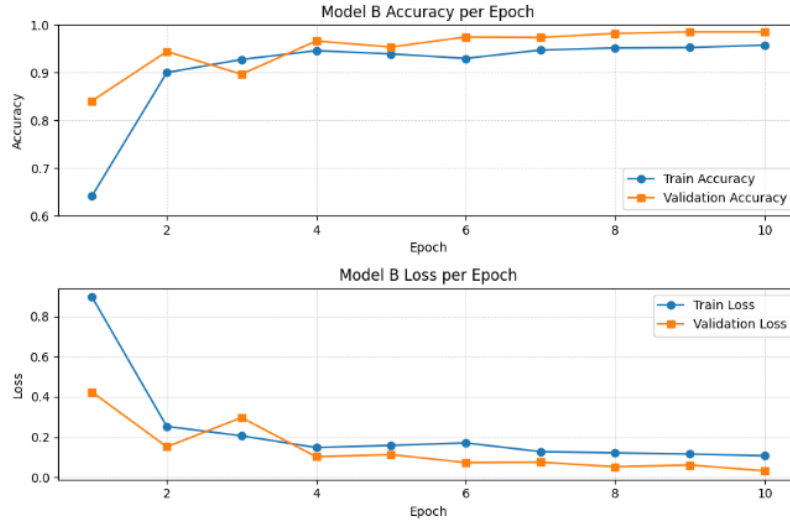


Figure 10: Training and validation accuracy/loss curves for ConvNeXt-Tiny (seed 42).

5.8 Discussion

The proposed dual-backbone ensemble achieves strong performance on the internal test set. The complementary architectures capture diverse features: fine micro-textures in EfficientNet-B0 and large-receptive-field patterns in ConvNeXt-Tiny. Logit-level fusion (Eq. 7) mitigates overconfident errors by combining two independent decision boundaries before the final normalisation.

The high F1 scores for healthy classes confirm that the model does not frequently misdiagnose healthy leaves as diseased. The remaining errors concentrate on visually similar diseases (GLS vs. NLB), suggesting that additional spectral features or higher-resolution imaging may further improve discrimination. The ECE of 0.009 before temperature scaling represents a well-calibrated model; temperature scaling ($T=1.356$) provides an additional 22.8 % ECE reduction recommended for deployment.

Computationally, the ensemble requires ≈ 4.87 GFLOPs per inference—manageable on modest GPU hardware for offline batch inference in agricultural monitoring.

5.9 Comparison with Lightweight CNN Approaches

Table 9 compares the proposed framework with recent lightweight CNN approaches. These comparisons are approximate and non-commensurate: different studies use different datasets, splits and evaluation protocols. No state-of-the-art claim is made.

5.10 Benchmark Comparison with Standard CNN Models

Table 10 summarizes metrics for common CNN backbones in plant-disease recognition. Cross-study comparisons are approximate.

6 Limitations and Future Work

External/Cross-Dataset Generalization The most critical limitation is that all results are reported on a held-out split of the same data distribution used for training,

Table 9: Comparison with lightweight CNN models for multi-class maize or potato leaf-disease recognition. Caution: different studies use different datasets and protocols; direct numerical comparison is approximate only. No SOTA claim is made.

Study (Year)	Dataset / Classes	Model (params)	Acc (%)	Notes
Ours	7,005 maize + potato; 7 classes	EfficientNet-B0 (≈ 4 M) + ConvNeXt-Tiny (≈ 27.8 M)	99.24 (3-seed mean)	Logit-level fusion; internal split only; no external validation.
LFMNet [23] (2023)	Maize; 9 classes	LFMNet (0.88 M)	94.12	Multi-scale lightweight CNN.
Weighted-loss CNN [24] (2024)	Maize; 7 classes	EfficientNet-B0 + weighted loss	94.80	5,155-image dataset.
RegNetY-400MF [9] (2026)	Fruit Fly; 4 classes	RegNetY-400MF (8.3 M)	96.10	0.41 GFLOPs; pest detection in chili plantations.
SpudNet5-R3 [25] (2025)	Potato; 3 classes	Hybrid CNN	95.56–99.07	Uses CLAHE and Grad-CAM; evaluates on PlantVillage and PLD datasets.
Hybrid CNN [26] (2025)	Same potato; 7 classes	Hybrid SqueezeNet blocks	+ 94.70	Shallow CNN; preprint.
EffNet-B0+SVM [27] (2023)	Cross-crop; 7 classes (~ 7 k img)	EfficientNet-B0 features + SVM	97.29	Cross-crop approach.
ConvNeXtSmall [28] (2024)	Potato; 7 classes	Custom ConvNeXtSmall	98.97	Different class set.

which is predominantly PlantVillage-derived. We do not claim that the model generalises to real-field conditions without external validation. The model has not been tested on PlantDoc [30] or any other independent field dataset, and prior literature shows that PlantVillage-trained models can drop from above 98 % to approximately 40–60 % accuracy on PlantDoc. All language suggesting “real-field readiness” has been removed or qualified. For future work we propose: (1) zero-shot transfer to PlantDoc maize/potato classes; (2) few-shot domain adaptation using 10 % of PlantDoc images; and (3) leave-one-source-out validation across the three data sources.

Data Deduplication Formal pHash exact-duplicate removal was performed prior to all experiments (Section 3.1.4). Six training images that were exact duplicates of test images were removed (Table 3). Four validation/test duplicates were identified and documented but not removed, as they do not affect training. Near-duplicate (Hamming distance > 0) removal was not performed; results may therefore be marginally optimistic if visually similar but non-identical images appear across splits.

Table 10: Benchmark comparison of popular CNN models for plant-disease recognition. Missing values indicate metrics not reported in the original source. Cross-study comparisons are approximate only; no SOTA claim is made.

Model / Study	Dataset (crops)	Acc (%)	F1 (%)	GFLOPs	Params (M)	Strategy
Ours (Ensemble)	Maize & Potato	99.24	98.66	≈ 4.87	≈ 31.8	Logit fusion ^a
ResNet50 (Haque et al. [3])	Maize / Potato	98.98	98.40	3.80	25.6	Transfer learning
EfficientNet-B0 (Tan and Le [7])	Maize	92.33	—	0.39	4.0	Single backbone
MobileNetV2 (Shoaib et al. [21])	Potato	93.20	92.80	0.30	3.4	Lightweight CNN
DenseNet121 (Luong [28])	Multi-crop	99.81	—	2.80	7.0	Dense connections
VGG16 (Sagar and Rajasekar [27])	Potato	91.00	90.50	15.5	138.0	Traditional CNN
^a Logit-level weighted average: $\mathbf{z}_{\text{final}} = 0.05 \mathbf{z}_{\text{Eff}} + 0.95 \mathbf{z}_{\text{Conv}}$, then $\sigma(\mathbf{z}_{\text{final}})$; see Eqs. 7–8. Multi-seed mean at $\alpha=0.30$. GFLOPs via <code>torchinfo</code> at 224×224 input; params include 7-class head replacement.						

Calibration With an ECE of 0.0091 (pre-scaling) and 0.0070 (post-scaling), the model is well calibrated by the standards of deep ensembles. Temperature scaling ($T=1.356$) applied to $\mathbf{z}_{\text{final}}$ is recommended before deployment [31].

Ablation Completeness The ablation study covers four EfficientNet-B0 configurations across a single seed. Multi-seed ablation across both backbones is deferred to future work. The contribution of backbone choice in the ensemble (e.g., replacing ConvNeXt-Tiny with a transformer variant) is also a promising direction.

Rare Classes The Healthy Potato class has only 24 test samples. Bootstrapped 95 % CIs are provided (Table 5) and explicitly capture this uncertainty. Expanding the test set for underrepresented classes is strongly recommended.

7 Conclusion

This paper presented a dual-backbone CNN ensemble for automatic diagnosis of maize and potato leaf diseases. A heterogeneous dataset of 7,005 images was curated from three sources; formal pHash deduplication removed six exact train/test duplicate images, yielding a clean training set of 5,091 images. EfficientNet-B0 and ConvNeXt-Tiny were integrated via logit-level weighted fusion with $\alpha=0.05$ selected by a validation grid search, and trained under three independent random seeds. The ensemble achieved mean test accuracy of 0.9924 ± 0.0013 and mean macro-F1 of 0.9866 ± 0.0021 across seeds. An

expanded ablation study showed that Mixup augmentation provides the greatest single-backbone improvement ($F1 = 0.9882$). Calibration analysis yielded $ECE=0.0091$ before and 0.0070 after post-hoc temperature scaling ($T=1.356$), confirming a well-calibrated ensemble; temperature scaling is recommended for deployment. Bootstrapped 95 % confidence intervals confirm robustness across all seven classes. All AUC values reach at least 0.9990 (macro 0.9998), verified at eight decimal-place precision with no probability clipping. We explicitly acknowledge that results are reported on an internal test split and that real-field generalisation remains an open research question. Future work will conduct external validation on PlantDoc, perform extended multi-seed ablations, and explore transformer-based backbone variants.

Acknowledgements

We thank the anonymous reviewers for their constructive and thorough feedback, which substantially improved this manuscript. We acknowledge the open-source communities behind PyTorch and `timm`. Data sources are acknowledged in Section 3.

References

- [1] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272–279, 2019. doi: 10.1016/j.compag.2018.03.032.
- [2] S. Albahli and M. Masood, "Efficient attention-based CNN network (EANet) for multi-class maize crop disease classification," *Frontiers in Plant Science*, vol. 13, 2022. doi: 10.3389/fpls.2022.1003152.
- [3] M. A. Haque et al., "Deep learning-based approach for identification of diseases of maize crop," *Scientific Reports*, vol. 12, no. 1, p. 6334, 2022. doi: 10.1038/s41598-022-10140-z.
- [4] L. Liu et al., "A multi-scale feature fusion neural network for multi-class disease classification on the maize leaf images," *Heliyon*, vol. 10, no. 7, p. e28264, 2024. doi: 10.1016/j.heliyon.2024.e28264.
- [5] A. Kolesnikov et al., "Big Transfer (BiT): General Visual Representation Learning," *arXiv preprint*, 2020. <https://arxiv.org/abs/1912.11370>.
- [6] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, 2021. <https://arxiv.org/abs/2010.11929>.
- [7] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv preprint*, 2020. <https://arxiv.org/abs/1905.11946>.
- [8] Z. Liu et al., "A ConvNet for the 2020s," *arXiv preprint*, 2022. <https://arxiv.org/abs/2201.03545>.
- [9] S. Rahman, A. Indrawati, M. Zen, B. Zealtiel, and S. S. Tanjung, "Enhanced RegNetY-400MF for Fruit Fly Species Classification: Fine-Tuning Strategies and Data Balancing for Improved Accuracy," *Jurnal RESTI (Rekayasa*

- Sistem Dan Teknologi Informasi*), vol. 9, no. 6, pp. 1463–1473, 2026. doi: 10.29207/resti.v9i6.6973.
- [10] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv preprint*, 2019. <https://arxiv.org/abs/1711.05101>.
- [11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *arXiv preprint*, 2018. <https://arxiv.org/abs/1710.09412>.
- [12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *arXiv preprint*, 2018. <https://arxiv.org/abs/1708.02002>.
- [13] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," *arXiv preprint*, 2017. <https://arxiv.org/abs/1608.03983>.
- [14] R. Müller, S. Kornblith, and G. Hinton, "When Does Label Smoothing Help?" *arXiv preprint*, 2020. <https://arxiv.org/abs/1906.02629>.
- [15] E. D. Cubuk et al., "AutoAugment: Learning Augmentation Policies from Data," *arXiv preprint*, 2019. <https://arxiv.org/abs/1805.09501>.
- [16] J. Chen, J. Chen, D. Zhang, Y. Sun, and Y. Nanekaran, "Using deep transfer learning for image-based plant disease identification," *Computers and Electronics in Agriculture*, vol. 173, p. 105393, 2020. doi: 10.1016/j.compag.2020.105393.
- [17] 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.
- [18] J. G. A. Barbedo, "Plant disease identification from individual lesions and spots using deep learning," *Biosystems Engineering*, vol. 180, pp. 96–107, 2019. doi: 10.1016/j.biosystemseng.2019.02.002.
- [19] M. Shoaib et al., "An advanced deep learning models-based plant disease detection: A review of recent research," *Frontiers in Plant Science*, vol. 14, 2023. doi: 10.3389/fpls.2023.1158933.
- [20] I. Pacal et al., "A systematic review of deep learning techniques for plant diseases," *Artificial Intelligence Review*, vol. 57, no. 11, p. 304, 2024. doi: 10.1007/s10462-024-10944-7.
- [21] M. Shoaib, A. Sadeghi-Niaraki, F. Ali, I. Hussain, and S. Khalid, "Leveraging deep learning for plant disease and pest detection: a comprehensive review and future directions," *Frontiers in Plant Science*, vol. 16, 2025. doi: 10.3389/fpls.2025.1538163.
- [22] S. Yang et al., "Classification and localization of maize leaf spot disease based on weakly supervised learning," *Frontiers in Plant Science*, vol. 14, 2023. doi: 10.3389/fpls.2023.1128399.
- [23] J. Hu, X. Jiang, J. Gao, and X. Yu, "LFMNet: a lightweight model for identifying leaf diseases of maize with high similarity," *Frontiers in Plant Science*, vol. 15, 2024. doi: 10.3389/fpls.2024.1368697.

- [24] K. Ahadian et al., "Maize disease classification using transfer learning and convolutional neural network with weighted loss," *Heliyon*, vol. 10, no. 21, p. e39569, 2024. doi: 10.1016/j.heliyon.2024.e39569.
- [25] S. Roy et al., "SPUDNET5-R3: A Lightweight Hybrid and Explainable CNN Model for Potato Leaf Blight Detection," 2025. doi: 10.21203/rs.3.rs-7377242/v1.
- [26] T. Sultana and M. Reza, "Identification of Potato Leaf Diseases Using Hybrid Convolution Neural Network with Support Vector Machine," in *Communications in Computer and Information Science*, Springer, pp. 350–361, 2023. doi: 10.1007/978-3-031-35641-4_28.
- [27] G. Sangar and V. Rajasekar, "Optimized classification of potato leaf disease using EfficientNet-LITE and KE-SVM in diverse environments," *Frontiers in Plant Science*, vol. 16, 2025. doi: 10.3389/fpls.2025.1499909.
- [28] H. H. Luong, "Improving Potato Diseases Classification Based on Custom ConvNeXtSmall and Combine with the Explanation Model," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, 2024. doi: 10.14569/IJACSA.2024.01504121.
- [29] M. M. Jawed et al., "A hybrid deep learning framework using convolutional and transformer models for robust plant disease classification," *Scientific Reports*, vol. 16, no. 1, p. 9704, 2026. doi: 10.1038/s41598-026-38209-z.
- [30] D. Singh et al., "PlantDoc," in *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, New York, NY, USA, pp. 249–253, 2020. doi: 10.1145/3371158.3371196.
- [31] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," *arXiv preprint*, 2017. <https://arxiv.org/abs/1706.04599>.