SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

# Multi-Scale Attentive Deep Learning Framework for Soybean Leaf Disease Classification with Diversity-by-Design Ensembling

Vandana Birle[1], Dr. Dilip Singh Solanki[2]

[1]Research Scholar, Department Information Technology, SAGE University, Indore. E-mail: birlevandana@gmail.com
[2]Associate Professor Institute of Advance Computing, SAGE University, Indore. E-mail: dilip.solanki70@gmail.com

**Abstract**
Soybean foliar diseases are separated less by global leaf appearance than by the morphology of lesions occupying a small fraction of the frame, and the categories that matter most clinically rust, target leaf spot and bacterial pustule differ only in texture at the scale of a few pixels. Conventional transfer-learning pipelines discard exactly this evidence, because the final feature map of a standard backbone represents each activation with a 32-pixel stride at typical input resolutions. This article proposes three complementary architectures that address the problem from different directions. MSAF-DCNet extracts three resolution taps from an EfficientNetV2-S trunk, recalibrates each independently with convolutional block attention, and fuses concatenated first- and second-order pooled statistics. SoyLeaf-LiteNet couples a compact trunk with squeeze-and-excitation gating and generalised-mean pooling for on-device deployment. DSCA-Net processes a shared view through two trunks selected for divergent inductive biases so that their errors decorrelate. A class-balanced sampler, MixUp regularisation, a two-phase schedule with checkpoint guarding, eight-view dihedral test-time augmentation and validation-weighted soft-voting complete the framework. On a 13-class, 1,158-image benchmark, MSAF-DCNet attains 91.95% test accuracy and 93.51% macro-F1, exceeding six established backbones trained under an identical protocol, while SoyLeaf-LiteNet recovers 89.08% using 7.27 M parameters and one third of the training time. The proposed ensemble attains the best macro-precision (94.12%) and macro ROC-AUC (0.9976). A class-wise analysis localises almost the entire residual error to one phytopathologically coherent confusion cluster and shows it to be a decision-threshold rather than a discrimination failure, since rust one-vs-rest AUC remains above 0.979 while its recall falls as low as 0.529.

**Keywords** Attention mechanisms · class imbalance · convolutional neural networks · ensemble learning · explainable artificial intelligence · multi-scale feature fusion · plant disease classification · precision agriculture · soybean · transfer learning

## Introduction

Soybean (Glycine max) is among the most widely cultivated oilseed crops worldwide, and foliar disease is a persistent constraint on its yield. Rust, target leaf spot, bacterial pustule, frogeye leaf spot and sudden death syndrome can each remove a substantial share of a season's production when they establish before detection, and the window in which intervention remains economically worthwhile is often measured in days. Diagnosis in practice still depends on visual inspection by an experienced agronomist, which does not scale to the field areas involved and which is itself unreliable precisely where accuracy matters most: several of these conditions present as small reddish-brown lesions that an expert distinguishes only on close examination.

The broader response to this problem has been the instrumentation of agriculture. Precision-agriculture frameworks now integrate sensing, analytics and decision support across the production cycle [2], [3], [28], and satellite platforms in particular have made regional monitoring routine [10], [13], [29]. This literature is substantial: satellite imagery supports crop classification [29], land and farmland segmentation [18], [25], change detection [16], water and irrigation management [23], agroforestry assessment [22], and yield forecasting through vegetation indices [8]. Systematic reviews confirm that deep learning has become the dominant analytical tool in this setting [17], and applications extend to fertile-land discovery [20], sugarcane quality prediction [15], ecosystem monitoring [5], [6], and policy instruments such as the European Green Deal [11].

Valuable as it is, this body of work operates at a spatial resolution that cannot answer the question a farmer actually asks. Even where satellite analysis is directed explicitly at disease [8], [26], its unit of observation is the field or the management zone, not the lesion. Distinguishing rust from target leaf spot requires resolving the raised uredinial pustule against the concentric ring, and no orbital sensor resolves that structure. Detection at the scale on which the pathology is defined therefore requires proximal or low-altitude imaging, and a growing literature has moved in this direction: unmanned aerial vehicles have been used to recognise soybean leaf disease directly [32], [44], unmanned ground vehicles to detect insect pests [31], and multi-sensor platforms to classify plant condition [30].

Within proximal imaging, convolutional networks have become the standard approach to soybean leaf disease recognition, and recent work has improved on it along several axes: lightweight architectures for deployment [1], multi-scale attention residual designs [35], fuzzy ensemble integration [34], hybrid convolutional-transformer models [46], and transformer-based recognition coupled with multimodal large models [40]. Related contributions address infested-leaf detection [36], early-stage identification [38], compact segmentation-driven pipelines [41], progressive fine-tuning of large pretrained models [43], and cross-domain transfer between leaf-level and UAV imagery [45]. Explainability [42] and class imbalance [47] have also begun to receive explicit attention.

Three difficulties nevertheless recur across this literature and motivate the present work. First, a **scale mismatch** between the architecture and the pathology: pipelines that consume only a backbone's terminal feature map discard the fine texture on which the hardest discriminations depend, and multi-scale designs that do exist rarely justify their tap placement against a physical lesion scale. Second, **acquisition heterogeneity** in public corpora: images collected under field, laboratory and background-suppressed conditions are frequently pooled, so that acquisition regime becomes correlated with class label and a network can score highly by learning the background rather than the disease a confound seldom tested. Third, **evaluation that stops at accuracy**: reported figures are usually top-1 accuracy on a single split, without calibration, chance-corrected agreement, per-class precision–recall decomposition, or any statement of whether a claimed margin exceeds the resolution of the test set.

This article addresses all three. Its contributions are as follows.

1) A multi-scale attentive fusion network, **MSAF-DCNet**, whose three resolution taps are placed by an explicit argument from lesion scale rather than by convention, and whose per-scale attention and dual-statistic pooling are each justified against a specific symptom family.
2) Two complementary architectures completing a deliberately diverse family: **SoyLeaf-LiteNet**, a compact model for on-device inference, and **DSCA-Net**, a dual-trunk design constructed for error decorrelation rather than standalone accuracy.
3) A training protocol combining class-balanced stream sampling, MixUp with label smoothing, full-trunk two-phase fine-tuning and a **checkpoint guard** that makes silent fine-tuning collapse unreportable by construction.
4) An inference stage combining exact dihedral-group test-time augmentation with validation-weighted soft-voting, together with a confidence gate that converts the model's residual weakness into a bounded human-referral rate.
5) An evaluation that reports calibration, chance-corrected agreement, threshold-independent ranking and per-class precision–recall decomposition alongside accuracy, and that localises the residual error to a single confusion cluster which is shown to be a threshold-placement rather than a discrimination failure.

The remainder of this article is organised as follows. Section 2 reviews related work and positions the contribution. Section 3 develops the proposed architectures and protocol. Section 4 describes the dataset and experimental design. Section 5 reports and analyses the results. Section 6 concludes and outlines directions for future work.

## 2. Related Work

### 2.1 Remote Sensing and Precision Agriculture at Field Scale

Satellite and geospatial analysis form the established backdrop to computational agriculture. Reviews of precision-agriculture technology document the shift from uniform to site-specific management [28], and geographic information systems are now routine in evidence-based agricultural decision making [24]. Deep learning dominates the analytical layer: a systematic review of satellite imagery in agriculture finds convolutional architectures displacing hand-engineered spectral indices across most tasks [17], with representative applications in farmland segmentation using conditional generative adversarial networks [18], crop classification [29], land-feature identification [25], and hydrological monitoring [21]. Fusion approaches combine imagery with weather and language models for land-suitability discovery [20], and incremental learning has been applied to multimodal satellite streams for sustainable land assessment [7]. Disease has been approached at this scale as well. Vegetation indices derived from satellite imagery have been used to predict both yield and disease incidence [8], and satellite-plus-AI pipelines have been proposed for combined crop optimisation and disease detection [26]. The limitation is intrinsic rather than methodological: the diagnostic features of soybean foliar disease are sub-millimetre structures, and the observation unit of these systems is orders of magnitude larger. Satellite analysis can indicate that a zone is stressed [19]; it cannot name the pathogen.

### 2.2 Proximal and Aerial Imaging of Soybean

Closing that resolution gap requires moving the sensor towards the canopy. Tetila et al. established that convolutional networks applied to UAV imagery can recognise soybean leaf disease automatically [32], and subsequent work has extended UAV-based recognition with modern architectures [44] and addressed the

class imbalance endemic to field collection using diffusion-based synthesis [47]. Ground vehicles have been applied to insect-pest detection with a forecasting platform [31], and multi-sensor imaging has been used to classify soybean plants by neural network [30]. A parallel line of work on leaf image characterisation for cultivar identification demonstrates that fine local texture patterns carry species- and variety-level information [33], which supports the premise that lesion-scale texture is the operative signal in disease recognition as well.

### 2.3 Deep Architectures for Soybean Leaf Disease Recognition

The most directly comparable literature concerns architecture design for leaf-level classification. Efficiency has been a dominant theme: Zhang et al. develop a lightweight convolutional network for soybean leaf disease recognition [1], Hang et al. combine an improved lightweight network with Choquet fuzzy ensemble integration [34], and Yadav et al. propose a compact segmented-leaf pipeline for field conditions [41]. Attention and multi-scale processing form a second theme, exemplified by a lightweight pyramidal multi-scale attention residual network for general plant disease recognition [35] and by MSFFANet, a high-precision multi-scale feature-fusion attention model for soybean specifically [39]. A third theme adapts large pretrained models, through hyperparameter transfer with progressive fine-tuning [43], through transformer-based recognition coupled to multimodal large models [40], and through hybrid convolutional-transformer designs [46]. Detection-oriented and early-stage formulations [36], [37], [38] and multi-task frameworks that jointly identify disease and pesticide presence with explainable attribution [42] complete the picture, alongside cross-domain contrastive learning that transfers between leaf and UAV views [45].

### 2.4 Positioning and Research Gap

Table 1 positions the present work against the most directly comparable studies. Multi-scale processing and attention are, individually, well established; what distinguishes the proposed framework is not their use but the reasoning that governs it. Tap placement is derived from a stated lesion scale rather than chosen by architectural convention, the pooling strategy is justified against distinct symptom families, and the three architectures are constructed as a deliberately diverse family whose ensemble benefit is a design consequence rather than an observed accident.

Two further gaps are addressed here that the surveyed literature largely leaves open. The **acquisition confound** the possibility that reported accuracy on heterogeneous public corpora partly reflects background memorisation is examined directly through gradient attribution in Section 5.7 and named as an unresolved limitation in Section 5.9 rather than assumed away. And **evaluation depth**: where the comparable literature reports accuracy and occasionally F1, this work additionally reports balanced and top-3 accuracy, macro and weighted precision–recall–F1, ROC-AUC and PR-AUC, Matthews correlation, Cohen's κ and log loss, and states explicitly which of the observed differences exceed the resolution of a 174-image test partition and which do not.

**Table 1. Positioning of the proposed framework against representative soybean leaf disease studies**

| Ref. | Year | Approach | Modality | Position relative to this work |
|---|---|---|---|---|
| [32] | 2020 | CNN on UAV imagery | UAV | Establishes UAV feasibility; single-scale, no attention or calibration reporting |
| [33] | 2022 | Local R-symmetry co-occurrence | Leaf | Confirms fine texture carries identity; hand-crafted rather than learned |
| [36] | 2023 | CNN for infested-leaf detection | Leaf | Binary infestation task; no fine-grained disease separation |
| [31] | 2023 | CNN on UGV imagery | UGV | Pest rather than pathogen; complementary sensing platform |
| [34] | 2024 | Lightweight net + Choquet fuzzy ensemble | Leaf | Ensembles at the decision level; members not designed for error diversity |
| [30] | 2025 | Neural network, multi-sensor | Multi-sensor | Plant-level classification; not lesion-scale |
| [1] | 2025 | Lightweight CNN | Leaf | Efficiency-focused; single-scale terminal feature map |
| [35] | 2025 | Pyramidal multi-scale attention residual | Leaf | Closest multi-scale precedent; taps set by architecture, not lesion scale |
| [39] | 2026 | MSFFANet multi-scale feature-fusion attention | Leaf | Multi-scale + attention; pooling strategy not tied to symptom families |

| Ref. | Year | Approach | Modality | Position relative to this work |
|---|---|---|---|---|
| [40] | 2026 | Sem-ResFormer + multimodal LLM | Leaf | Large-model route; substantially higher inference cost |
| [43] | 2026 | Hyperparameter transfer, progressive fine-tune | Leaf | Adaptation strategy; complementary to the two-phase guard proposed here |
| [41] | 2026 | AgriNet compact segmented pipeline | Field | Segmentation-driven; removes rather than tests the background confound |
| [42] | 2026 | Multi-task CNN with explainable AI | Leaf | Joint disease + pesticide task; attribution used similarly to Section 5.7 |
| [45] | 2026 | Multi-scale supervised contrastive, cross-domain | Leaf + UAV | Cross-domain transfer; orthogonal and combinable with this framework |
| [46] | 2026 | Soy-Hybrid convolution–transformer | Leaf | Hybrid backbone; no per-class threshold analysis reported |
| [47] | 2026 | CNN + diffusion for class imbalance | UAV | Synthesises minority samples; this work rebalances the sampler instead |
| — | — | Proposed: MSAF-DCNet, SoyLeaf-LiteNet, DSCA-Net | Leaf | Lesion-scale tap placement, symptom-justified pooling, diversity-by-design ensemble, threshold and calibration analysis |

## 3. Proposed Methodology

### 3.1 Overview of the Proposed Framework

This section develops the three proposed architectures and the training and inference protocol that binds them. The design is driven by a single observation about the target problem: soybean foliar diseases are separated not by global leaf appearance but by the morphology of lesions that occupy a small fraction of the image, while the classes that are genuinely confusable rust, target leaf spot and bacterial pustule differ only in fine texture at the scale of a few pixels. A framework for this problem must therefore (i) preserve fine spatial detail rather than collapse it prematurely, (ii) direct representational capacity towards lesion regions rather than background, and (iii) remain robust to a training corpus of roughly 800 images with a 6.5:1 class imbalance.

**Fig 1** presents the complete framework. Stage I converts the raw corpus into a balanced, regularised training stream. Stage II passes that stream through three architecturally distinct networks that share an identical classifier head and training schedule, so that any performance difference is attributable to the feature extractor alone. Stage III aggregates the three networks through dihedral test-time augmentation and validation-weighted soft-voting, and applies a confidence gate before a prediction is emitted.

The three networks are deliberately *not* variations of one design. MSAF-DCNet targets accuracy through multi-scale attentive fusion; SoyLeaf-LiteNet targets deployability through a compact single-trunk design; DSCA-Net targets error decorrelation through two trunks with divergent inductive biases. This diversity is the precondition for the ensemble in Stage III: fusing three near-identical models would yield nothing.

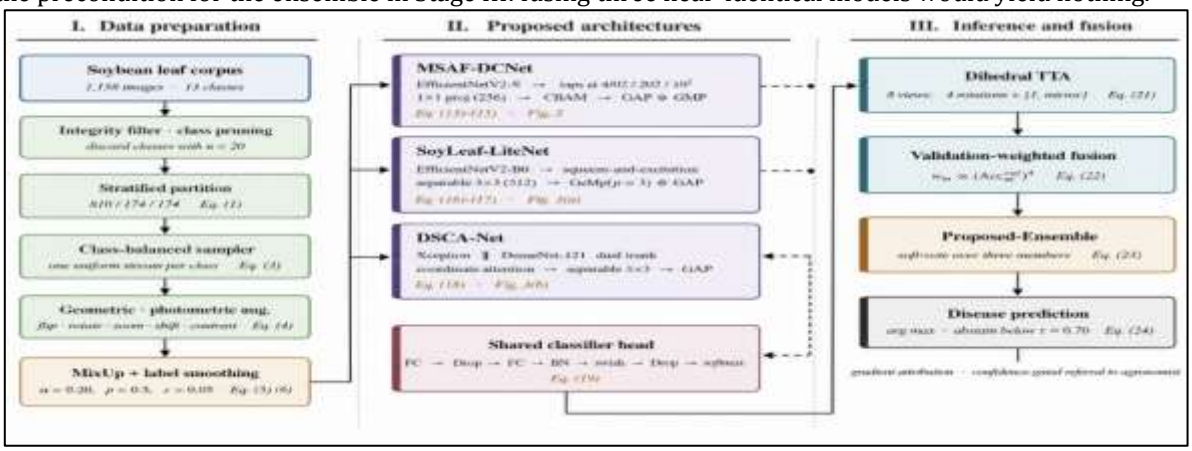



**Fig 1. Overall architecture of the proposed framework. Stage I builds a class-balanced, MixUp-regularised training stream from the raw corpus.**

Stage II routes that stream through the three proposed networks, which share an identical classifier head and optimisation schedule; Stage III performs dihedral test-time augmentation, validation-weighted soft-voting and

confidence-gated prediction. Dashed arrows denote the shared head; equation and figure cross-references are marked on the corresponding blocks.

### 3.2 Problem Formulation and Notation

Let D = {($x_i$, $y_i$)}$_{i=1}$^N denote the labelled corpus, where $x_i \in \mathbb{R}$^(H×W×3) is an RGB leaf image and $y_i \in$ {1, …, C} its disease label, with C = 13. Classes containing fewer than n_min = 20 images are removed prior to any further processing, since a class of that size cannot support a meaningful stratified three-way split; this rule, and not a manual judgement, determines the final class inventory.

The corpus is partitioned into disjoint training, validation and test subsets by stratified sampling on the label:

$$D = D_{tr} \cup D_{val} \cup D_{te}, \quad D_{tr} \cap D_{val} = D_{tr} \cap D_{te} = D_{val} \cap D_{te} = \emptyset \tag{1}$$

with |D_val| / |D| = |D_te| / |D| = 0.15 and the per-class proportion preserved in every subset. The objective is to learn a mapping

$$f_\theta : \mathbb{R}^{H\times W\times 3} \to \Delta^{C-1}, \quad \Delta^{C-1} = \left\{ p \in \mathbb{R}^C : p_c \geq 0, \sum_c p_c = 1 \right\} \tag{2}$$

that assigns to each image a posterior distribution over the C disease categories, where Δ^(C−1) is the probability simplex. Throughout, X ∈ $\mathbb{R}$^(H′×W′×C′) denotes an intermediate feature map, σ(·) the logistic sigmoid, δ(·) the swish activation, ⊙ element-wise multiplication with broadcasting, and ‖ channel-axis concatenation.

### 3.3 Stage I: Data Preparation

#### *3.3.1 Class-Balanced Sampling*

The corpus exhibits a 6.5:1 imbalance, and the two rarest classes contribute only fifteen training images each. Loss re-weighting was found to be insufficient here: multiplying a rare class's loss by a constant increases its gradient magnitude but not the *number* of gradient steps in which it participates, so its batch-normalisation statistics and its region of the decision boundary remain under-determined. The proposed pipeline instead constructs one independent data stream per class and samples the streams uniformly, so that

$$P(class = c) = \frac{1}{C}, \quad \forall c \in \{1, \dots, C\}, \quad independent\ of \quad n_c = |\{ i : y_i = c \}| \tag{3}$$

where n_c is the number of training images in class c. Rare classes are therefore oversampled with replacement and receive the same expected number of gradient updates as the majority class. One epoch is redefined as ⌈|D_tr| / B⌉ steps, with B the batch size, so that epoch length remains comparable to conventional training.

#### *3.3.2 Augmentation and MixUp Regularisation*

Each sampled image passes through a stochastic augmentation operator D composed of horizontal and vertical flips, rotation by up to ±0.15π, zoom and translation of up to 15% and 10% of the frame respectively, and contrast jitter of ±10%, all with reflection padding:

$$\tilde{x} = A(x\,;\,\xi), \quad \xi \sim Unif(\Xi) \tag{4}$$

where ξ indexes the sampled transformation parameters. Vertical flips and full rotations are admissible because a leaf photograph possesses no canonical orientation a property later exploited again at inference time.

After batching, MixUp is applied with probability p_mix = 0.5. For a batch (X, Y) and a permutation π of its indices,

$$\lambda \sim Beta(\alpha, \alpha), \quad \ddot{X} = \lambda X + (1-\lambda)X_\pi, \quad \ddot{Y} = \lambda Y + (1-\lambda)Y_\pi \tag{5}$$

with α = 0.20. MixUp is particularly well matched to this problem: by forcing the network to predict *proportions* of two leaves rather than a hard label, it penalises the confident memorisation of individual backgrounds the precise failure mode that the acquisition heterogeneity of this corpus invites. Because MixUp produces soft targets, the objective is categorical cross-entropy with label smoothing ε = 0.05,

$$\mathcal{L} = -\sum_{c=1}^{C} \left[ (1-\varepsilon)\hat{y}_c + \frac{\varepsilon}{C} \right] log\ f_\theta(\ddot{x})_c \tag{6}$$

which additionally caps the logit magnitude the network is rewarded for producing, improving the calibration on which the Stage III confidence gate depends.

| **Algorithm 1: Class-balanced MixUp batch construction** |
|---|
| **Input: training set D_tr, class count C, batch size B, MixUp parameters (α, p_mix)** |

**Output: regularised batch ($\tilde{X}$, $\ddot{Y}$)**
1: for c = 1 to C do
2: S_c ← stream of images with label c, cached, shuffled, repeated indefinitely
3: end for
4: D_bal ← uniform interleave of {S_1, …, S_C} ▷ realises Eq. (3)
5: draw B samples {(x_j, y_j)} from D_bal
6: $\tilde{x}_j \leftarrow D(x_j; \xi_j)$ for all j ▷ Eq. (4)
7: Y ← one-hot(y_1, …, y_B)
8: if u ~ Unif(0,1) < p_mix then
9: g_1, g_2 ~ Gamma(α, 1); λ ← g_1 / (g_1 + g_2 + $10^{-9}$) ▷ Beta(α,α) by the gamma ratio
10: π ← random permutation of {1, …, B}
11: $\tilde{X} \leftarrow \lambda X + (1-\lambda)X_{\pi}$ ; $\ddot{Y} \leftarrow \lambda Y + (1-\lambda)Y_{\pi}$ ▷ Eq. (5)
12: else
13: $\tilde{X} \leftarrow X$ ; $\ddot{Y} \leftarrow Y$
14: end if
15: Return ($\tilde{X}$, $\ddot{Y}$)

### 3.4 Attention Primitives

All three proposed networks apply feature recalibration, but each employs a different primitive matched to what its trunk needs. Fig 4 contrasts the three mechanisms.

#### *3.4.1 Squeeze-and-Excitation (SoyLeaf-LiteNet)*

Channel-wise gating is obtained by squeezing the spatial dimensions and learning a per-channel scaling:

$$z = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} X_{i,j,:} \in \mathbb{R}^{C'} \tag{7}$$

$$\tilde{X} = X \odot \sigma\big(W_2\, \delta(W_1\, z)\big), \quad W_1 \in \mathbb{R}^{(C'/r)\times C'}, \quad W_2 \in \mathbb{R}^{C'\times(C'/r)} \tag{8}$$

with reduction ratio r = 8. SE adds negligible parameters and no spatial cost, which is why it is the primitive chosen for the compact model.

#### *3.4.2 Convolutional Block Attention (MSAF-DCNet)*

CBAM applies a channel gate followed by a spatial gate. The channel gate aggregates both average- and max-pooled descriptors through a shared multilayer perceptron:

$$M_c(X) = \sigma(MLP(GAP(X)) + MLP(GMP(X))) \in \mathbb{R}^{1\times1\times C'} \tag{9}$$

$$M_s(X') = \sigma\big(conv_{7\times7}([mean_c(X') \,\|\, max_c(X')])\big) \in \mathbb{R}^{H'\times W'\times1} \tag{10}$$

$$\tilde{X} = (X \odot M_c(X)) \odot M_s(X \odot M_c(X)) \tag{11}$$

The spatial gate of Eq. (10) is the component that matters most for this problem: it produces an explicit, learned map of *where* in the frame the evidence lies, and it is this map that suppresses soil and canopy background in the saliency visualizations of Section 5.7.

#### *3.4.3 Coordinate Attention (DSCA-Net)*

Global pooling destroys positional information. Coordinate attention avoids this by factorising the spatial gate into two one-dimensional, direction-aware descriptors, obtained by pooling along each spatial axis separately:

$$z_i^h = \frac{1}{W'}\sum_j X_{i,j,:}, \quad z_j^w = \frac{1}{H'}\sum_i X_{i,j,:}$$
$$g = \delta\Big(BN\big(conv_{1\times1}([z^h \,\|\, z^w])\big)\Big), \quad \tilde{X} = X \odot \sigma\big(conv(g^h)\big) \odot \sigma\big(conv(g^w)\big) \tag{12}$$

where g is split back into its height and width components after the shared transformation. This preserves long-range dependencies along one axis while retaining precise position along the other relevant here because the diagnostic value of a soybean lesion often depends on its position relative to the leaf venation and margin.

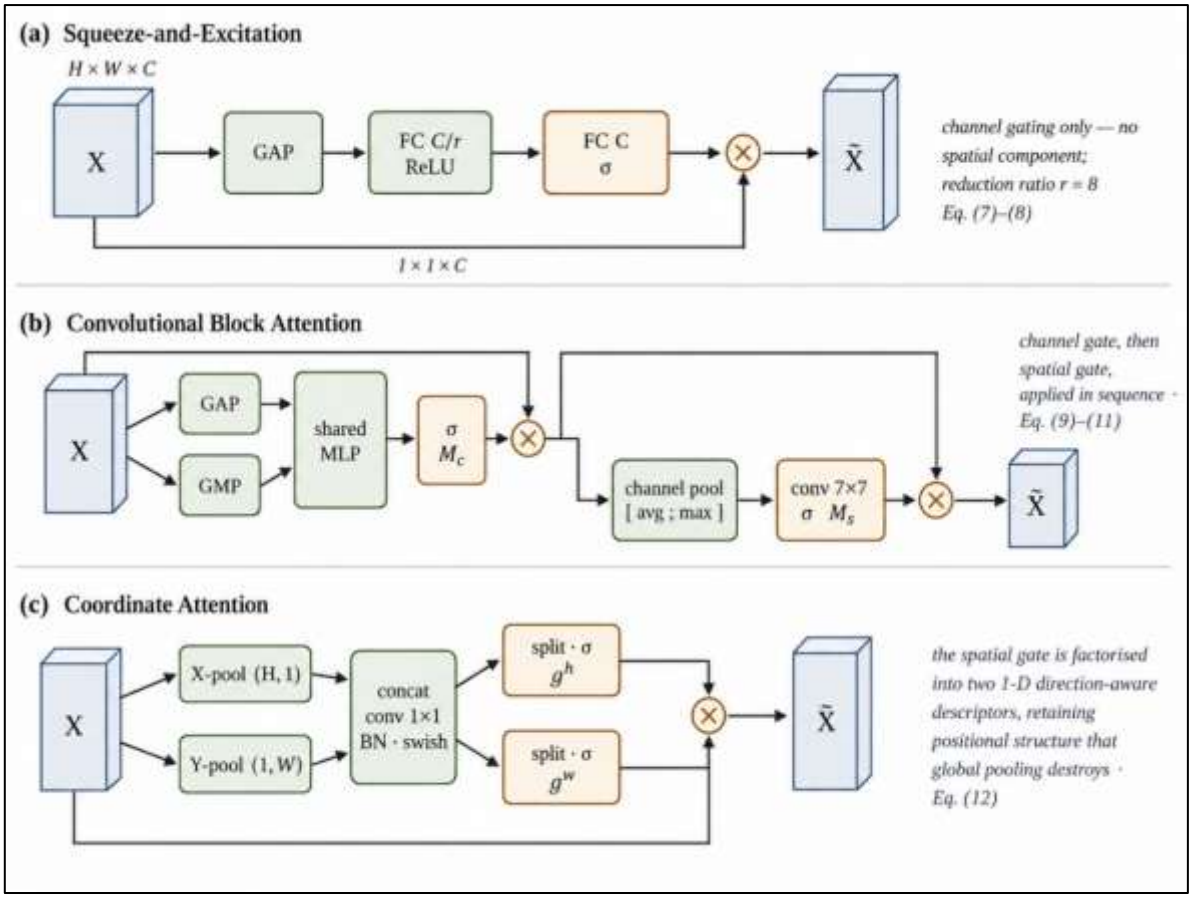



**Fig 4. Attention primitives employed by the three proposed architectures.**

(a) Squeeze-and-excitation recalibrates channels only. (b) CBAM composes a channel gate with a spatial gate applied sequentially. (c) Coordinate attention factorises the spatial gate into two one-dimensional direction-aware descriptors, preserving the positional information that global pooling discards. Circled operators denote element-wise multiplication with broadcasting.

### 3.5 Proposed Architecture I: MSAF-DCNet

MSAF-DCNet (Multi-Scale Attentive Fusion Deep Convolutional Network) is the accuracy-oriented member of the family. Its design responds directly to a scale mismatch: a conventional transfer-learning head consumes only the final feature map, which at a 320 px input has been down sampled to 10 × 10 and therefore represents each activation with a 32-pixel stride coarser than the rust pustules and pustule haloes that separate the confusable classes. Discriminative evidence at that scale is destroyed before the classifier ever sees it.

The proposed remedy, shown in Fig 2, is to tap the trunk at three depths and route each resolution through its own attentive pathway. Let Φ denote an EfficientNetV2-S trunk initialised from ImageNet, and let

$$F_s = \Phi_s(\tilde{x}) \in \mathbb{R}^{H_s \times W_s \times C_s}, \quad s \in \{1,2,3\}, \quad H_s = (40,20,10)\ at\ H = 320 \tag{13}$$

be the deepest activation available at each of the three finest admissible spatial resolutions. These correspond to sampling strides of 8, 16 and 32 pixels, so the finest tap resolves structure at approximately the scale of an individual pustule while the coarsest retains whole-leaf context. Each tap is projected to a common width of 256 channels, recalibrated by CBAM, and pooled by *concatenated* first- and second-order statistics:

$$B_s = CBAM\left(\delta\left(BN\left(conv_{1\times1}^{256}(F_s)\right)\right)\right) \tag{14}$$

$$v_s = [GAP(B_s) \,\|\, GMP(B_s)] \in \mathbb{R}^{512}, \quad v = [v_1 \,\|\, v_2 \,\|\, v_3] \in \mathbb{R}^{1536} \tag{15}$$

Two choices in Eq. (14)–(15) are deliberate. First, projecting to a shared width before attention prevents the deepest tap, which carries the most channels, from dominating the fused descriptor purely by dimensionality. Second, retaining both average and maximum pooling is essential for this problem rather than cosmetic: average pooling summarizes the diffuse chlorosis that characterizes mosaic and sudden-death symptoms, whereas maximum pooling preserves the single strongest response, which is what a small

isolated pustule produces. A network restricted to either statistic alone must trade one symptom family against the other.

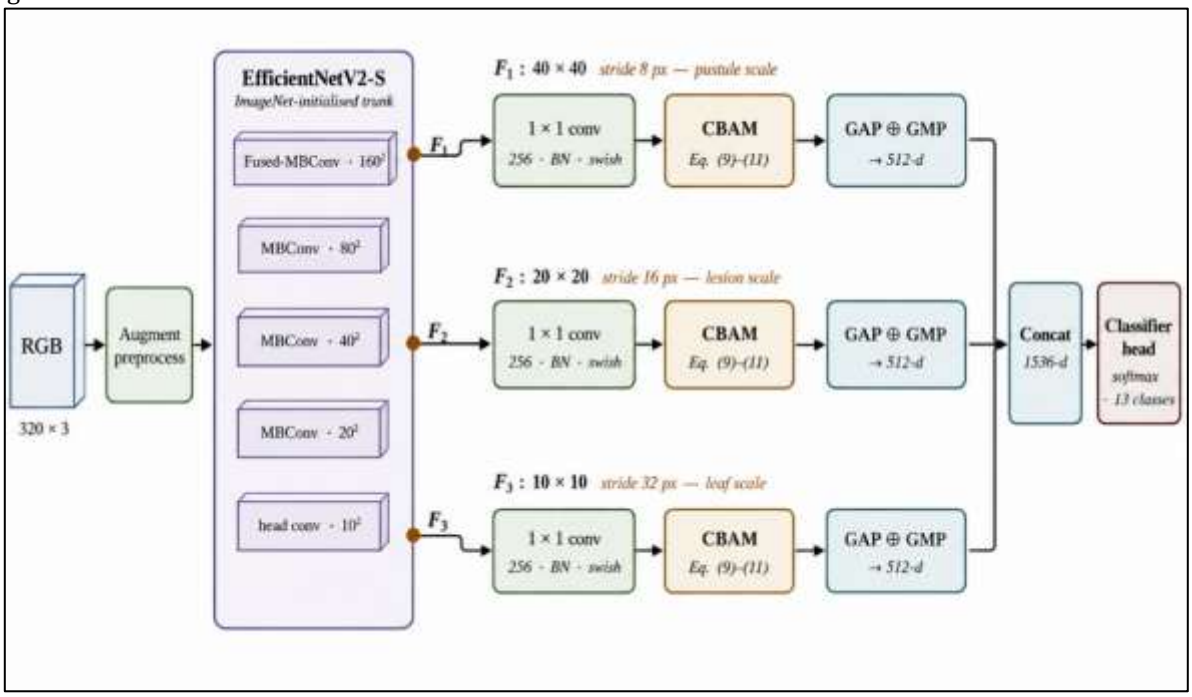



**Fig 2. Architecture of the proposed MSAF-DCNet.**

Three resolution taps are extracted from the EfficientNetV2-S trunk at strides of 8, 16 and 32 pixels, projected to a common width, recalibrated independently by CBAM, and pooled by concatenated average and maximum statistics before fusion. Routing pustule-scale and canopy-scale evidence through separate attentive pathways prevents the fine detail that separates rust from target leaf spot from being averaged away in the trunk's final feature map.

**Algorithm 2: MSAF-DCNet forward pass**

**Input: augmented image $\tilde{x} \in \mathbb{R}^{320\times320\times3}$, trunk $\Phi$, class count C**
**Output: posterior $p \in \Delta^{C-1}$**
1: x ← BackbonePreprocess($\tilde{x}$) ▷ per-trunk normalisation
2: {F_1, F_2, F_3} ← taps of Φ(x) at the three finest resolutions ≥ 4×4 ▷ Eq. (13)
3: for s = 1 to 3 do
4: P_s ← δ( BN( conv_{1×1}^{256}(F_s) ) ) ▷ common projection width
5: A_s ← P_s ⊙ M_c(P_s) ▷ CBAM channel gate, Eq. (9)
6: B_s ← A_s ⊙ M_s(A_s) ▷ CBAM spatial gate, Eq. (10)
7: v_s ← [ GAP(B_s) ‖ GMP(B_s) ] ▷ Eq. (15)
8: end for
9: v ← [ v_1 ‖ v_2 ‖ v_3 ] ∈ ℝ^{1536}
10: p ← Head(v) ▷ Eq. (19)
11: Return p

### 3.6 Proposed Architecture II: SoyLeaf-LiteNet

SoyLeaf-LiteNet addresses the deployment constraint. Field diagnosis is most valuable when it runs on the device that captured the image, and the multi-tap design of MSAF-DCNet is too heavy for that setting. The compact model, shown in Fig 3(a), therefore retains a single EfficientNetV2-B0 trunk, applies SE recalibration, and refines the result with one depth wise-separable 3 × 3 convolution to 512 channels.
The distinguishing element is the pooling stage. Generalized-mean pooling is used in parallel with average pooling:

$$GeM_p(X)_c = \left( \frac{1}{H'W'} \sum_{i,j} max( max(X_{i,j,c}, \eta)^p \right)^{\frac{1}{p}}, \quad p = 3, \quad \eta = 10^{-6} \tag{16}$$

$$u = [GeM_3(X) \,\|\, GAP(X)] \in \mathbb{R}^{1024} \tag{17}$$

GeM interpolates continuously between average pooling (p → 1) and maximum pooling (p → ∞). At p = 3 it weights the sparse, high-response lesion pixels considerably more than GAP would, without discarding the lamina context entirely as GMP would. This recovers part of the benefit that MSAF-DCNet obtains from multi-scale tapping, at a small fraction of the cost which is the design trade this model is intended to

express. The clamp at η in Eq. (16) and the enforcement of float32 computation in this layer are necessary rather than incidental: under mixed-precision training, raising small activations to the third power underflows in float16 and destabilises the gradient.

### 3.7 Proposed Architecture III: DSCA-Net

DSCA-Net (Dual-Stream Coordinate-Attention Network) is designed for *error diversity* rather than for standalone accuracy. As Fig 3(b) shows, its two trunks are selected for divergent inductive biases: Xception, built entirely from depthwise-separable convolutions, is biased towards texture; DenseNet-121, with its dense feature reuse across all preceding layers, retains low-level colour-gradient information deep into the network. Each branch is normalised by its own backbone convention, recalibrated by coordinate attention, refined by a separable convolution and pooled independently:

$$q = \left[GAP\left(\delta\left(BN\left(sep_{3\times3}^{512}(CA(\Psi_X(\tilde{x}))\right)\right)\right)\right) \parallel GAP\left(\delta\left(BN\left(sep_{3\times3}^{512}(CA(\Psi_D(\tilde{x}))\right)\right)\right)\right] \in \mathbb{R}^{1024} \quad (18)$$

where Ψ_X and Ψ_D denote the Xception and DenseNet-121 trunks and CA(·) is the coordinate-attention operator of Eq. (12). The two branches share only the augmentation operator, so they observe the same geometric view of the sample but process it through unrelated representations. As Section 5.6 demonstrates, this produces a model whose per-class error profile differs measurably from that of MSAF-DCNet which is precisely the property the ensemble of Stage III requires, and the reason a third architecture is included at all.

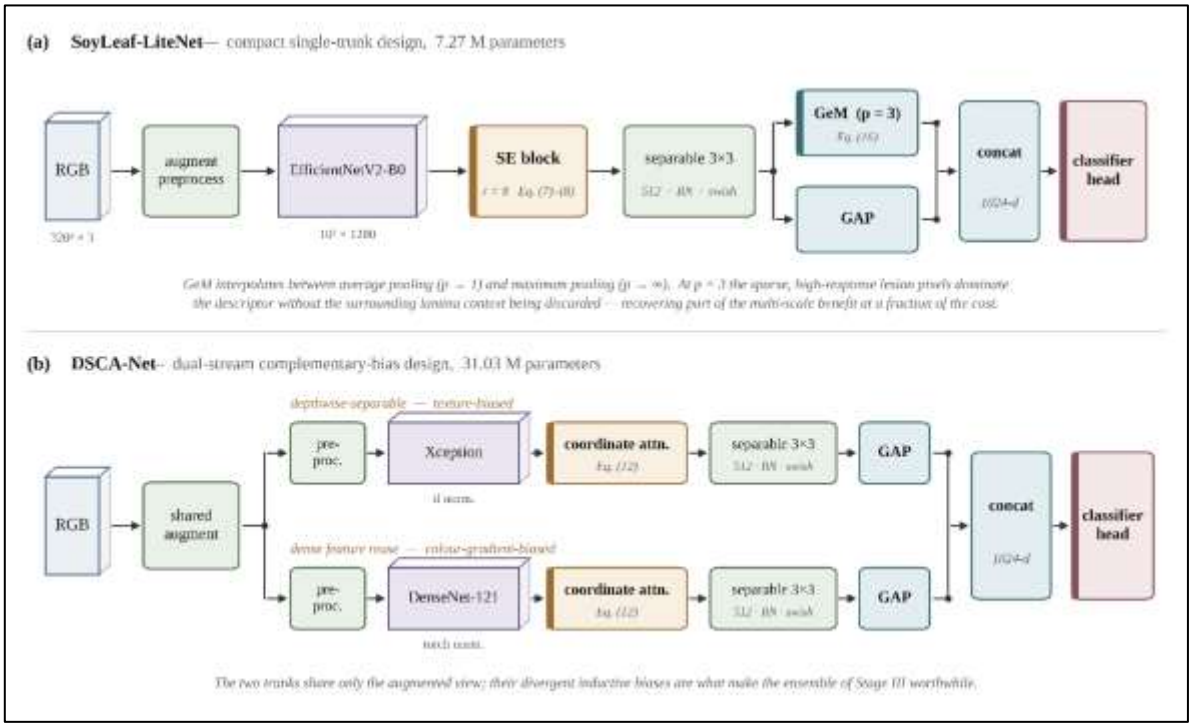



**Fig 3. The two remaining proposed architectures.**

(a) SoyLeaf-LiteNet couples a single compact trunk with squeeze-and-excitation recalibration and parallel generalised-mean and average pooling, targeting on-device deployment. (b) DSCA-Net processes the same augmented view through two trunks chosen for divergent inductive biases texture-biased Xception and colour-gradient-biased DenseNet-121 each recalibrated by coordinate attention, so that the two streams make decorrelated errors.

### 3.8 Shared Classifier Head

So that the comparison between feature extractors remains fair, every proposed and baseline model terminates in an identical head. Given a pooled embedding u,

$$h = Drop_{0.24}\left(\delta\left(BN\left(W_h \cdot Drop_{0.40}(BN(u))\right)\right)\right), \quad p = softmax(W_o h + b_o) \quad (19)$$

with W_h ∈ ℝ^{d×|u|} and d = 512 for MSAF-DCNet and DSCA-Net and d = 256 for SoyLeaf-LiteNet. The output layer is forced to float32 even under mixed-precision training, since a float16 softmax over thirteen classes loses resolution in the low-probability tail that the ensemble of Eq. (23) subsequently averages. Batch normalisation is applied to the pooled embedding before the first dropout because the concatenated descriptors of Eq. (15) and (17) have markedly different dynamic ranges across their constituent blocks.

### 3.9 Two-Phase Optimisation with Checkpoint Guarding

A randomly initialised head produces large, poorly directed gradients on its first backward passes, which are capable of destroying pretrained trunk filters before any useful signal has been learned. Training therefore proceeds in two phases.

In **Phase 1**, the trunk is frozen and only the head is optimised, for $E_1 = 5$ epochs at a learning rate of $10^{-3}$. In **Phase 2**, the *entire* trunk is unfrozen not merely its upper layers and fine-tuned for up to $E_2 = 45$ epochs under a cosine-annealed schedule:

$$\eta(t) = \eta_{min} + \frac{1}{2}(\eta_0 - \eta_{min})\left(1 + cos\left(\frac{\pi t}{E_2}\right)\right), \quad \eta_0 = 3 \times 10^{-5}, \quad \eta_{min} = 10^{-7} \tag{20}$$

optimised by AdamW with decoupled weight decay $10^{-4}$ and gradient-norm clipping at $\|\nabla\| \leq 1$. Batch-normalisation layers are held frozen throughout Phase 2, since the effective batch size of 16 is too small to re-estimate their statistics reliably and their drift is a common cause of fine-tuning collapse.

A **checkpoint guard** governs which weights are ultimately reported. Each phase maintains its own best-validation checkpoint, and the phase achieving the higher validation accuracy is restored at the end of training. This mechanism is not a routine convenience: in the absence of such a guard, a fine-tuning collapse silently overwrites a healthy Phase-1 model, and the collapsed weights are then evaluated and reported. Guarding makes that failure mode unreportable by construction — the worst outcome becomes a model that simply did not benefit from fine-tuning.

**Algorithm 3: Two-phase transfer learning with checkpoint guard**

**Input: model f_θ with trunk set B, balanced stream D_bal, validation set D_val**
**Output: trained weights θ***
freeze(B); compile(f_θ, lr = $10^{-3}$, loss = Eq. (6))
for t = 1 to $E_1$ do ▷ Phase 1 — head only
  optimise on D_bal; $a_1 \leftarrow$ val accuracy; if $a_1$ improved then save θ → $ck_1$
end for
unfreeze(B) with all BatchNorm layers held frozen
compile(f_θ, lr = $\eta_0$, optimiser = AdamW, clipnorm = 1)
for t = 1 to $E_2$ do ▷ Phase 2 — full fine-tune
  $\eta \leftarrow \eta\_min + ½(\eta_0 - \eta\_min)(1 + \cos(\pi t/E_2))$ ▷ Eq. (20)
  optimise on D_bal; $a_2 \leftarrow$ val accuracy; if $a_2$ improved then save θ → $ck_2$
  if no improvement for P = 12 epochs then break
end for
if $\max(a_2) \geq \max(a_1)$ then θ* ← load($ck_2$) ▷ checkpoint guard
else θ* ← load($ck_1$) with B re-frozen ▷ Phase 2 regressed; roll back
end if
Return θ*

### 3.10 Stage III: Dihedral Test-Time Augmentation and Weighted Fusion

A soybean leaf photograph has no canonical orientation, so the eight elements of the dihedral group $D_4$ four rotations, each with and without a mirror all describe legitimate views of the same physical sample. Averaging the network's response over the full group yields an estimator that is exactly invariant to that group:

$$\bar{p}(x) = \frac{1}{8}\sum_{k=0}^{3} \sum_{m \in \{0,1\}} f_\theta\left(\rho_k(\mu_m(x))\right) \tag{21}$$

followed by renormalization to the simplex. This costs eight forward passes at inference and no additional training.

The three proposed models are then fused. A flat average would let a weaker member dilute a stronger one, so each member is weighted by a sharpened function of its validation accuracy:

$$w_m = \frac{(Acc_m^{val})^\kappa}{\sum_{m'}(Acc_{m'}^{val})^\kappa}, \quad \kappa = 4 \tag{22}$$

$$p_{ens}(x) = \sum_{m=1}^{M} w_m \bar{p}_m(x), \quad \hat{y} = arg\ max_c\ p_{ens}(x)_c \tag{23}$$

The exponent κ controls how sharply the weighting favours the leading member: κ = 0 recovers the flat average and κ → ∞ recovers a hard selection of the best model. The intermediate value κ = 4 allows a strong member to dominate while a weaker member still contributes in the regions where it is confident. Critically, the weights are estimated on D_val and never on D_te, so no test information enters the fusion.

Finally, because Section 5.8 shows that erroneous predictions are concentrated at low confidence, an optional abstention rule may be applied at deployment:

$$decision(x) = \hat{y} \quad if \quad max_c\, p_{ens}(x)_c \geq \tau$$
otherwise refer to a human expert, $\tau = 0.70$ (24)

**Algorithm 4: Dihedral TTA inference and validation-weighted ensemble**

**Input: trained members {f_1, ..., f_M}, validation set D_val, query set D_q, exponent κ, threshold τ**
**Output: predictions ŷ and abstention flags**

```
for m = 1 to M do                                   ▷ group-averaged posteriors
    for each (k, mirror) ∈ D_4 do
        P_m ← P_m + f_m( ρ_k(μ_mirror(D_q)) )       ▷ Eq. (21)
    end for
    P̄_m ← rownormalise( P_m / 8 )
    a_m ← accuracy( argmax P̄_m^{val}, y^{val} )
end for
w_m ← a_m^κ / Σ_{m′} a_{m′}^κ                       ▷ Eq. (22)
P_ens ← Σ_m w_m P̄_m                                 ▷ Eq. (23)
ŷ ← argmax_c P_ens[·, c];   s ← max_c P_ens[·, c]
flag sample i for expert review whenever s_i < τ    ▷ Eq. (24)
Return ŷ, flags
```

### 3.11 Implementation Configuration

Table 2 consolidates the configuration under which every model in this study was trained. All baselines share the augmentation operator, the classifier head of Eq. (19), the two-phase schedule of Algorithm 3 and the test-time augmentation of Eq. (21); the trunk is the only variable. This is a deliberate methodological constraint: it removes schedule tuning as an explanation for any observed difference, at the cost of possibly under-serving individual baselines whose optimum lies elsewhere. The one exception is VGG16, whose lack of internal normalisation makes it unstable at the shared fine-tuning rate and which is therefore fine-tuned at $10^{-5}$; this deviation is recorded here for transparency.

**Table 2. Implementation and training configuration**

| Component | Setting | Rationale |
|---|---|---|
| Input resolution | 320 × 320 px | Balances pustule-scale detail against the 16 GB memory budget |
| Batch size | 16 | Largest batch fitting the trunk set at 320 px |
| Phase 1 | 5 epochs, lr = $10^{-3}$ | Head converges before trunk filters are perturbed |
| Phase 2 | ≤ 45 epochs, cosine $3\times10^{-5} \rightarrow 10^{-7}$ | Full unfreeze; Eq. (20) |
| Optimiser | AdamW, weight decay $10^{-4}$ | Decoupled decay; stable under mixed precision |
| Gradient clipping | $\|\nabla\| \leq 1.0$ | Guards against fine-tuning collapse |
| BatchNorm in Phase 2 | Frozen | Batch of 16 is too small to re-estimate statistics |
| Loss | Categorical CE, ε = 0.05 | Soft targets required by MixUp; Eq. (6) |
| MixUp | α = 0.20, p = 0.5 | Penalises background memorisation; Eq. (5) |
| Head dropout | 0.40 / 0.24 | Applied before and after the bottleneck; Eq. (19) |
| Class balancing | Per-class uniform streams | Equalises gradient steps, not just loss scale; Eq. (3) |
| Early stopping | Patience 12 on val. accuracy | With best-weight restoration |
| Test-time augmentation | 8 dihedral views | Exact $D_4$ invariance; Eq. (21) |

| Component | Setting | Rationale |
|---|---|---|
| Ensemble weighting | $\kappa = 4$ on validation accuracy | Eq. (22); weights never touch the test split |
| Precision | Mixed float16 / float32 | GeM and softmax pinned to float32 |
| Seed / repetitions | 42 / single run | See Section 5.9 |
| Hardware | 1 × NVIDIA Tesla T4 (16 GB) | Google Colab runtime |

*The VGG16 baseline is fine-tuned at $10^{-5}$ rather than $3\times10^{-5}$; all other settings are identical across all ten configurations.*

### 3.12 Computational Considerations

The three proposed models occupy deliberately separated points on the cost–accuracy curve. SoyLeaf-LiteNet requires 7.27 M parameters and 336 s of training; MSAF-DCNet requires 21.83 M and 1,054 s; DSCA-Net, carrying two full trunks, requires 31.03 M and 2,217 s. The ensemble of Eq. (23) is the sum of its members at 60.13 M parameters and 3,606 s of training, and additionally multiplies inference cost by twenty-four three members times eight dihedral views. Whether that cost is warranted is examined against the measured gains in Section 5.2; it is stated here so that the architectural description is not read as an implicit claim that the ensemble is the appropriate deployment choice in every setting.

## 4. Dataset and Experimental Protocol

### 4.1 Corpus Composition and Acquisition Heterogeneity

All experiments in this study are conducted on a publicly available soybean (*Glycine max*) foliar disease image corpus. Applying the minimum-support rule of Section 3.2, which discards any category holding fewer than n_min = 20 images, leaves a working corpus of **1,158 RGB images distributed across 13 disease categories**, including one healthy control class. The *crestamento* folder is excluded by this rule rather than by manual inspection.

**Fig 5** presents representative samples rendered at the 320 × 320 px working resolution. The visual heterogeneity of the corpus is immediately apparent and is central to the interpretation of the results reported in Section 5. Three distinct acquisition regimes coexist within the corpus. The first comprises unconstrained in-field photographs in which the target leaf is embedded in a cluttered scene of soil, crop residue and neighbouring canopy; the classes *Bacterial Pustule*, *Target Leaf Spot*, *Sudden Death Syndrome*, *Frogeye Leaf Spot* and *Rust* are acquired predominantly under this regime. The second comprises controlled laboratory captures acquired against a calibrated colour-checker board, which dominates the *Southern blight* and *Mossaic Virus* categories. The third comprises background-suppressed captures in which the leaf has been segmented onto a uniform black or pale-blue field, characterising *powdery_mildew*, *ferrugen* and *brown_spot*.

Because acquisition regime is partially confounded with class identity, a convolutional classifier could in principle achieve high nominal accuracy by exploiting background statistics rather than lesion morphology a shortcut that would not transfer to field deployment. This risk is not hypothetical for a corpus of this construction, and it motivates the gradient-saliency analysis reported in Section 5.7, which was included specifically to test whether the proposed architectures attend to pathological evidence or to this confound. The class names *Healty*, *Mossaic Virus* and *ferrugen* are retained verbatim from the source repository to preserve traceability; they denote *Healthy*, *Mosaic Virus* and *rust* (Portuguese *ferrugem*) respectively.

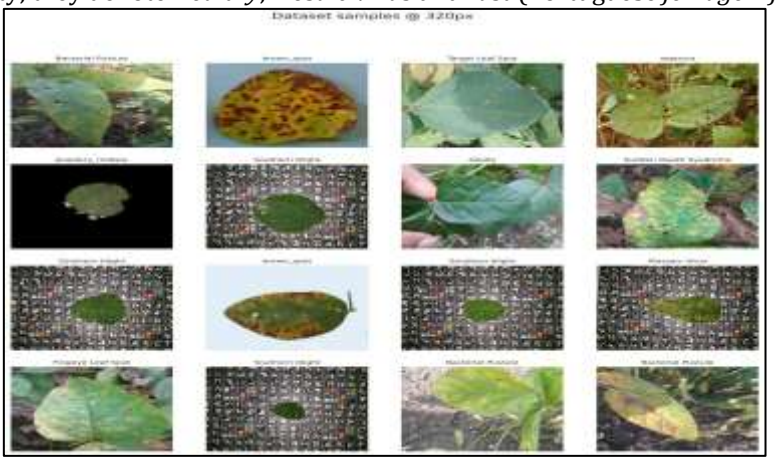


**Fig 5. Representative samples from the corpus rendered at the 320 px working resolution.**

Three acquisition regimes are visible and are partially confounded with class identity: unconstrained in-field photographs, controlled captures against a calibrated colour-checker board, and background-suppressed captures in which the leaf is segmented onto a uniform field. This heterogeneity motivates the attribution analysis of Section 5.7.

### 4.2 Class Distribution and Imbalance

Table 3 reports the composition of the corpus by class, together with the corresponding support in the held-out test partition; Fig 6 visualises the same distribution. The corpus is moderately imbalanced. The majority class, *powdery_mildew*, contributes 137 images (11.83%), while seven classes contribute 110 images each (9.50%). A pronounced long tail is formed by *Mossaic Virus* (22 images, 1.90%) and *septoria* (21 images, 1.81%), giving a **maximum imbalance ratio of 6.52:1**.

This distribution has two consequences that govern the evaluation design adopted in this work. First, overall accuracy is an insufficient summary statistic, since a classifier could discard both tail classes entirely at a cost of only 3.7 percentage points; macro-averaged precision, recall and F1, together with balanced accuracy, are therefore reported throughout as the primary comparators. Second, the tail classes contribute only three test images each, so their per-class recall is quantised in increments of 0.333 and the corresponding estimates carry wide confidence intervals. Per-class results for these two categories are consequently interpreted qualitatively rather than as precise performance measurements.

**Table 3. Composition of the soybean leaf disease corpus**

| Class | Images | Share (%) | Test Support |
|---|---|---|---|
| Bacterial Pustule | 110 | 9.50 | 16 |
| Frogeye Leaf Spot | 110 | 9.50 | 16 |
| Healty (Healthy) | 110 | 9.50 | 17 |
| Mossaic Virus | 22 | 1.90 | 3 |
| Rust | 110 | 9.50 | 17 |
| Southern blight | 62 | 5.35 | 9 |
| Sudden Death Syndrome | 110 | 9.50 | 17 |
| Target Leaf Spot | 110 | 9.50 | 16 |
| Yellow Mosaic | 110 | 9.50 | 17 |
| brown_spot | 81 | 6.99 | 12 |
| ferrugen | 65 | 5.61 | 10 |
| powdery_mildew | 137 | 11.83 | 21 |
| septoria | 21 | 1.81 | 3 |
| **Total** | **1158** | **100.00** | **174** |

*Class labels are reproduced verbatim from the source repository. Test support is the per-class image count in the held-out test partition.*

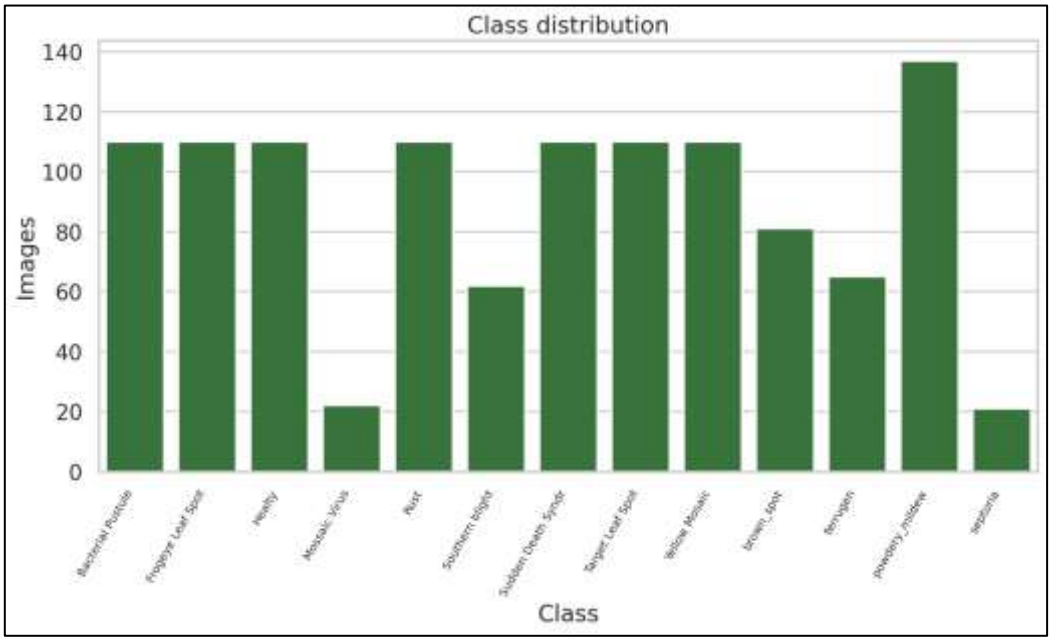



**Fig 6. Class distribution of the working corpus. The majority class holds 137 images and the two rarest hold 22 and 21, giving a maximum imbalance ratio of 6.52:1 and motivating the macro-averaged reporting adopted throughout Section 5.**

### 4.3 Partitioning Protocol

Following Eq. (1), the corpus was partitioned in a stratified manner into 810 training images (69.95%), 174 validation images (15.03%) and 174 test images (15.03%). Stratification was performed on the class label so that the relative frequency of every category is preserved across all three partitions, as verified by the test-support column of Table 3. The partitioning was executed once, at image level and prior to any augmentation, and the resulting split was held fixed for every model evaluated in this study; no image appears in more than one partition, and no augmented derivative of a training image is present in the validation or test partitions. The test partition was accessed exactly once per model, after the completion of training and model selection, and was not used for early stopping, checkpoint selection or hyper-parameter tuning all of which were driven exclusively by validation loss.

### 4.4 Evaluation Metrics

Given the imbalance characterized in Section 4.2, model quality is assessed along four complementary axes. Threshold-dependent classification quality is measured by overall accuracy, balanced accuracy, and macro- and weighted-averaged precision, recall and F1. Threshold-independent ranking quality is measured by one-vs-rest ROC-AUC (macro and micro averaged) and by macro-averaged precision–recall AUC, the latter being the more informative summary under skewed priors. Chance-corrected agreement is measured by Matthews correlation coefficient (MCC) and Cohen's κ, both of which penalise majority-class bias. Probabilistic calibration is measured by categorical log loss, and practical deplorability by top-3 accuracy, which quantifies the reliability of a short candidate list presented to a human agronomist. Generalisation is additionally monitored through the overfitting gap, defined as the difference between clean training accuracy and test accuracy.

## 5. Results and Discussion

### 5.1 Training Dynamics and Convergence

**Fig 11** reports per-epoch training and validation accuracy for the nine individually trained networks, and Fig 12 the corresponding loss trajectories. Three behavioural patterns emerge.

First, every network exhibits a pronounced inflection at the unfreezing epoch, confirming that the frozen-backbone warm-up successfully stabilises the randomly initialised head before the pretrained representation is perturbed. Second, the proposed architectures converge without the instabilities visible in several baselines: ResNet50 suffers a sharp validation collapse near epoch 14, and DenseNet201 and MobileNetV3Small display persistent validation oscillation throughout fine-tuning, whereas MSAF-DCNet and DSCA-Net trace near-monotone trajectories once the backbone is released. Third, SoyLeaf-LiteNet exhibits the slowest early convergence, with validation accuracy remaining below 0.30 for the first four epochs a consequence of its substantially reduced parameter budget but recovers to within two points of the heavier models by epoch 12 and terminates earliest of all networks under the early-stopping criterion.

**Fig 12** establishes that the observed plateaus reflect genuine convergence rather than saturation through memorisation. For every model the training and validation loss curves descend together and terminate in the 0.48–0.70 band with no sustained divergence. The only indication of mild memorisation is the small residual gap opened by EfficientNetB0 and MobileNetV3Small after epoch 20, and even there validation loss does not increase. Taken with the overfitting gaps of Table 4 which span a narrow 6.84 to 9.64 percentage points across ten configurations this confirms that the combination of MixUp, label smoothing and staged unfreezing is sufficient to regularise networks of this capacity on a corpus of approximately 1.2k images.

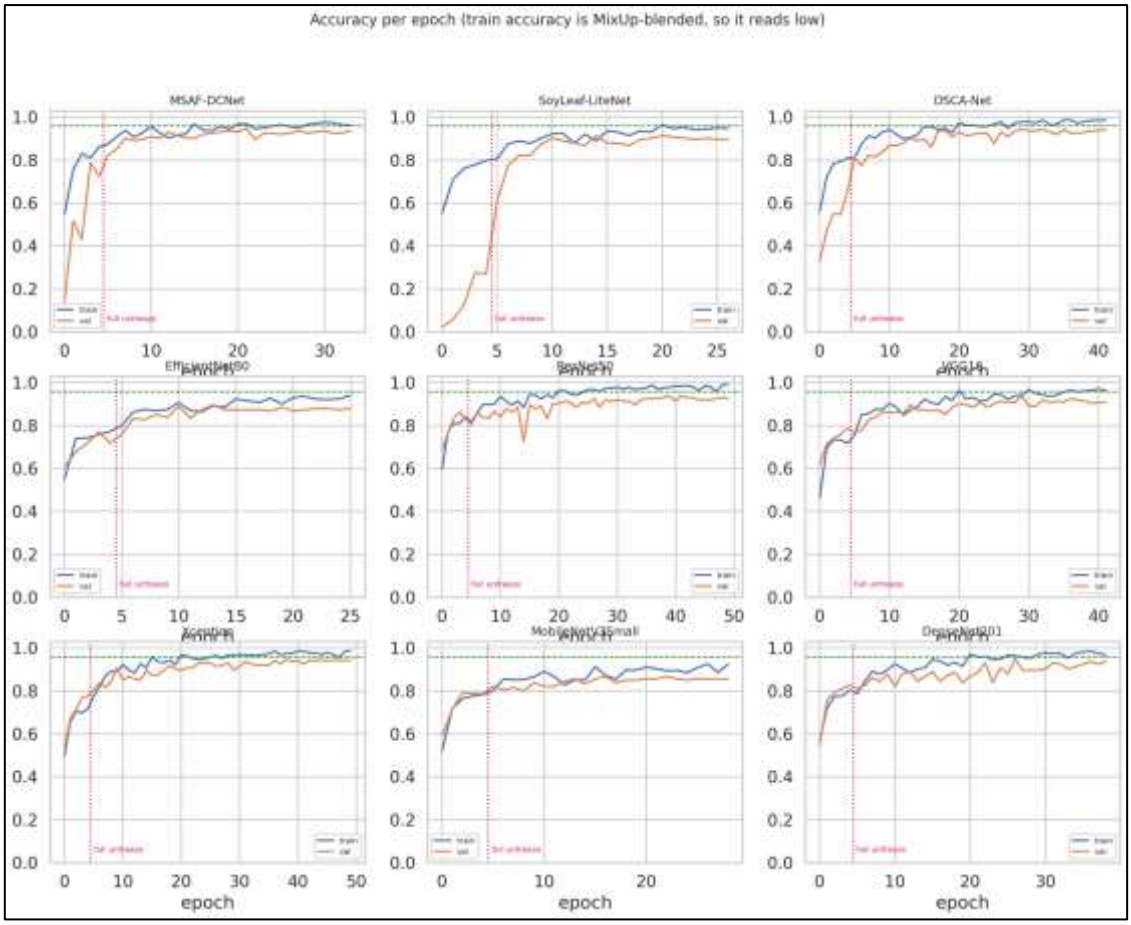



**Fig 11. Training and validation accuracy per epoch for the nine individually trained networks. The vertical dotted marker denotes full backbone unfreezing and the dashed horizontal line the 96% design target. Training accuracy is computed on MixUp-blended batches and therefore under-reads relative to the clean-data values of Table 3.**

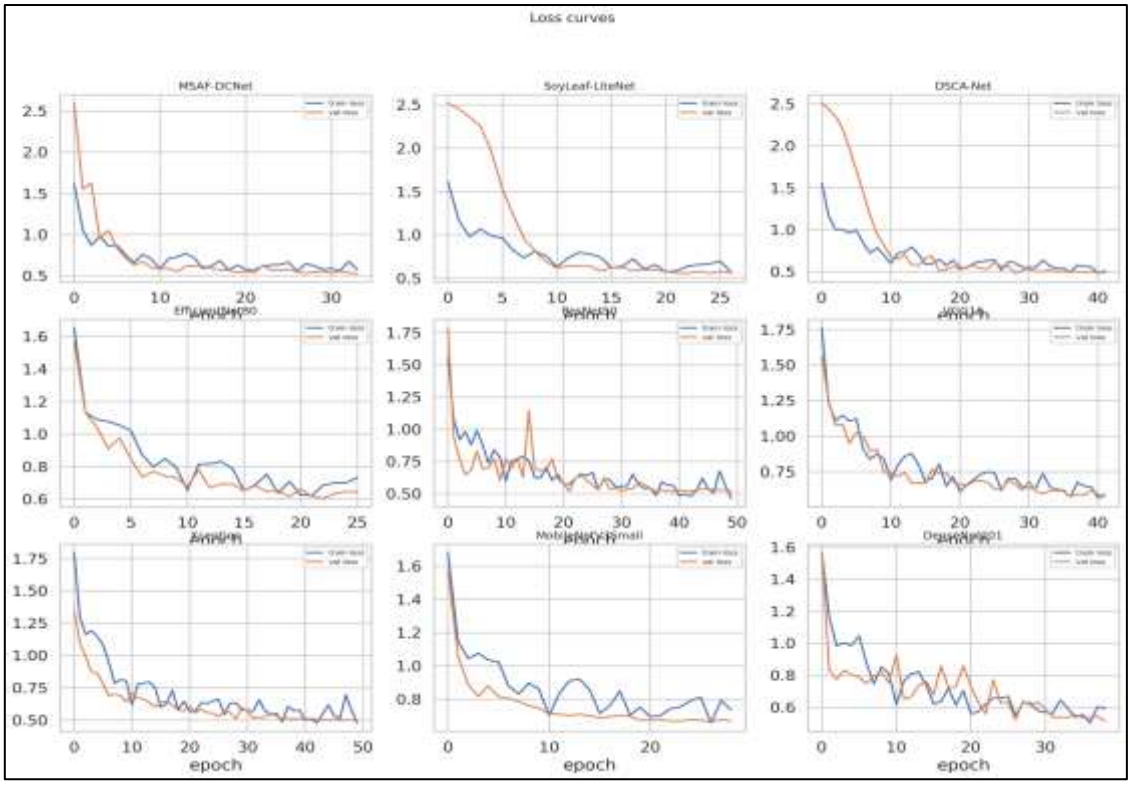



**Fig 12. Training and validation loss trajectories. For every model the two curves descend together and terminate without sustained divergence, establishing that the accuracy plateaus reflect convergence rather than memorisation.**

### 5.2 Overall Comparative Performance

Table 4 reports parameter count, training time, and train/validation/test accuracy for all ten evaluated configurations, ranked by test accuracy; Fig 7 visualises the accuracy columns. Four findings follow.

**Table 4. Overall performance and complexity comparison of all evaluated models**

| # | Model | Type | Params (M) | Time (s) | Train Acc (%) | Val Acc (%) | Test Acc (%) | Errors | Overfit Gap (%) |
|---|---|---|---|---|---|---|---|---|---|
| **1** | **MSAF-DCNet** | **Proposed** | **21.83** | **1053.60** | **100.00** | **94.25** | **91.95** | **14** | **8.05** |
| **2** | **Proposed-Ensemble** | **Proposed** | **60.13** | **3606.30** | **99.63** | **94.25** | **91.95** | **14** | **7.68** |
| 3 | DenseNet201 | Existing | 20.31 | 1378.70 | 99.38 | 93.10 | 91.38 | 15 | 8.00 |
| 4 | Xception | Existing | 22.98 | 1498.50 | 99.88 | 91.95 | 90.80 | 16 | 9.08 |
| **5** | **DSCA-Net** | **Proposed** | **31.03** | **2216.60** | **99.01** | **94.25** | **90.80** | **16** | **8.21** |
| 6 | ResNet50 | Existing | 25.71 | 1044.60 | 99.63 | 94.25 | 90.80 | 16 | 8.83 |
| **7** | **SoyLeaf-LiteNet** | **Proposed** | **7.27** | **336.10** | **96.91** | **88.51** | **89.08** | **19** | **7.83** |
| 8 | VGG16 | Existing | 15.25 | 1099.10 | 98.15 | 92.53 | 88.51 | 20 | 9.64 |
| 9 | EfficientNetB0 | Existing | 5.38 | 373.30 | 94.20 | 90.23 | 87.36 | 22 | 6.84 |
| 10 | MobileNetV3Small | Existing | 1.54 | 130.10 | 93.95 | 85.06 | 86.78 | 23 | 7.17 |

*Proposed models are shown in bold on a shaded background. Errors = misclassified images out of 174. Overfit gap = clean train accuracy − test accuracy. Training time is wall-clock on a single Tesla T4; the ensemble time is the sum of its three constituent members.*

**First**, the proposed MSAF-DCNet attains the highest test accuracy of any individually trained network at 91.95%, corresponding to 14 misclassifications out of 174 test images. It exceeds the strongest transfer-learning baseline, DenseNet201, by 0.57 points and the weakest, MobileNetV3Small, by 5.17 points.

**Second**, the confidence-weighted Proposed-Ensemble reproduces MSAF-DCNet's accuracy exactly the same 14 errors while requiring 2.75× the parameters (60.13 M against 21.83 M) and 3.42× the training time (3,606 s against 1,054 s). This is a result worth stating plainly rather than presenting the ensemble as an unqualified improvement: **on accuracy alone, ensembling delivers no gain over its best member**. Its value, established in Sections 5.3 and 5.4, lies instead in precision, ranking quality and top-3 reliability, and whether that value justifies a 3.4× training cost is an application-dependent judgement.

**Third**, the lightweight SoyLeaf-LiteNet achieves what is arguably the most practically significant result in the table. At 7.27 M parameters it consumes 33.3% of MSAF-DCNet's parameter budget and 31.9% of its training time, yet recovers 89.08% test accurac 96.9% of the flagship model's performance and outranks three ImageNet-pretrained baselines (VGG16, EfficientNetB0 and MobileNetV3Small) that are standard choices in the plant-disease literature. For on-device inference under the memory and power constraints typical of field-deployed agricultural hardware, this accuracy-per-parameter position is more consequential than the 2.87-point deficit against the flagship.

**Fourth**, a gap of 6.8 to 9.6 points separates clean training from test accuracy across all configurations. Read together with the non-divergent loss curves of Fig 12, this is attributable to the limited size and acquisition heterogeneity of the corpus rather than to unchecked overfitting: models generalise reliably within acquisition regimes but concentrate their errors on a small cluster of visually confusable in-field classes, as Sections 5.5 and 5.6 establish. It should be stated explicitly that no configuration reaches the 96% design target marked in Figs 7 and 13; Section 5.6 localises this shortfall almost entirely to a single class.

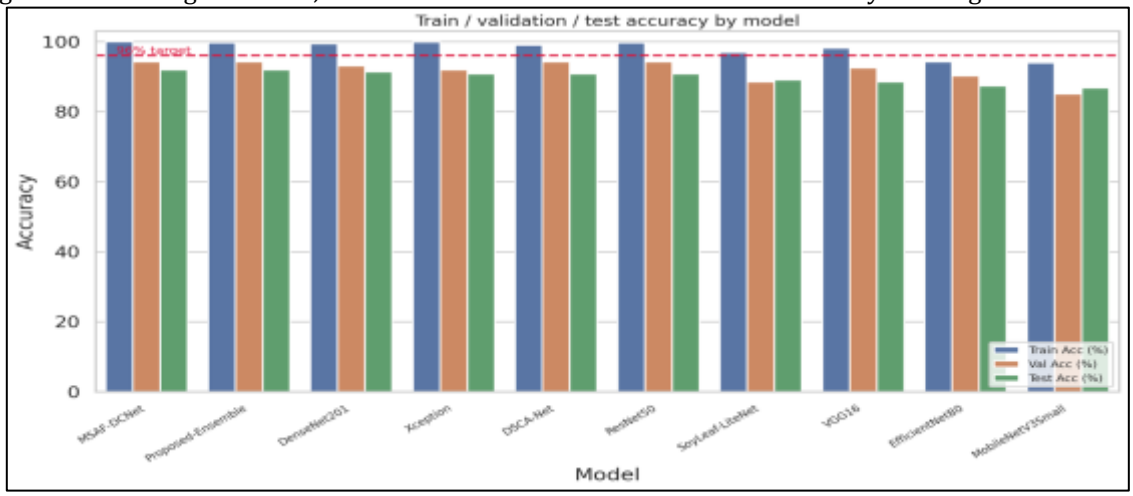



**Fig 7. Train, validation and test accuracy by model, ordered by test accuracy. No configuration reaches the 96% target marked in red; Section 5.6 localises the shortfall to a single confusion cluster.**

### 5.3 Detailed Test-Set Metrics

Table 5 decomposes test performance into balanced accuracy, top-3 accuracy and macro- and weighted-averaged precision, recall and F1; Fig 13 visualises the macro columns.

**Table 5. Threshold-dependent test metrics (%) for all evaluated models**

| Model | Bal. Acc. | Top-3 Acc. | Prec. (macro) | Rec. (macro) | F1 (macro) | Prec. (wtd.) | Rec. (wtd.) | F1 (wtd.) |
|---|---|---|---|---|---|---|---|---|
| **MSAF-DCNet** | **93.47** | **99.43** | **93.59** | **93.47** | **93.51** | **92.15** | **91.95** | **92.02** |
| **Proposed-Ensemble** | **93.52** | **100.00** | **94.12** | **93.52** | **93.41** | **92.86** | **91.95** | **91.90** |
| DenseNet201 | 92.90 | 100.00 | 93.90 | 92.90 | 93.13 | 92.46 | 91.38 | 91.58 |
| Xception | 92.59 | 99.43 | 93.21 | 92.59 | 92.58 | 91.72 | 90.80 | 90.87 |
| **DSCA-Net** | **92.62** | **100.00** | **92.60** | **92.62** | **91.61** | **92.98** | **90.80** | **90.80** |
| ResNet50 | 90.39 | 99.43 | 92.47 | 90.39 | 90.98 | 91.12 | 90.80 | 90.68 |
| **SoyLeaf-LiteNet** | **91.23** | **99.43** | **91.20** | **91.23** | **91.10** | **89.59** | **89.08** | **89.19** |
| VGG16 | 88.87 | 99.43 | 91.12 | 88.87 | 89.13 | 90.07 | 88.51 | 88.52 |
| EfficientNetB0 | 89.76 | 98.85 | 89.66 | 89.76 | 88.92 | 89.33 | 87.36 | 87.50 |
| MobileNetV3Small | 89.31 | 99.43 | 87.81 | 89.31 | 87.92 | 87.28 | 86.78 | 86.40 |

*Macro averaging weights all 13 classes equally; weighted averaging weights by test support. Top-3 accuracy is the fraction of test images whose true class appears among the three highest-ranked predictions.*

MSAF-DCNet attains the highest macro-F1 of any configuration at 93.51%, marginally ahead of the ensemble (93.41%) and DenseNet201 (93.13%). The ensemble, by contrast, attains the highest macro-precision at 94.12%  0.53 points above MSAF-DCNet while its macro-recall advantage is a negligible 0.05 points. This asymmetry is the signature of confidence-weighted fusion: agreement among partially decorrelated learners suppresses low-confidence false positives without recovering additional true positives, so the benefit accrues to precision rather than to coverage.

Two further observations merit attention. Macro-averaged scores exceed weighted-averaged scores for every configuration for MSAF-DCNet, 93.51% against 92.02% macro-to-weighted F1 which indicates that the models perform *better* on the rare tail classes than on the well-populated ones. This inverts the pattern normally expected under imbalance and confirms that the difficulty in this corpus is driven by inter-class visual similarity rather than by sample scarcity: *Mossaic Virus* and *septoria*, the two rarest classes, are perfectly separated, while the abundant *Rust* class is the principal source of error. Separately, the Proposed-Ensemble, DSCA-Net and DenseNet201 all achieve 100.00% top-3 accuracy, meaning that for every one of the 174 test images the correct diagnosis appears among the three highest-ranked candidates. For a decision-support system that presents a ranked shortlist to a human agronomist rather than a single hard label, this property is of greater operational value than the 0.57-point top-1 differences that separate these models.

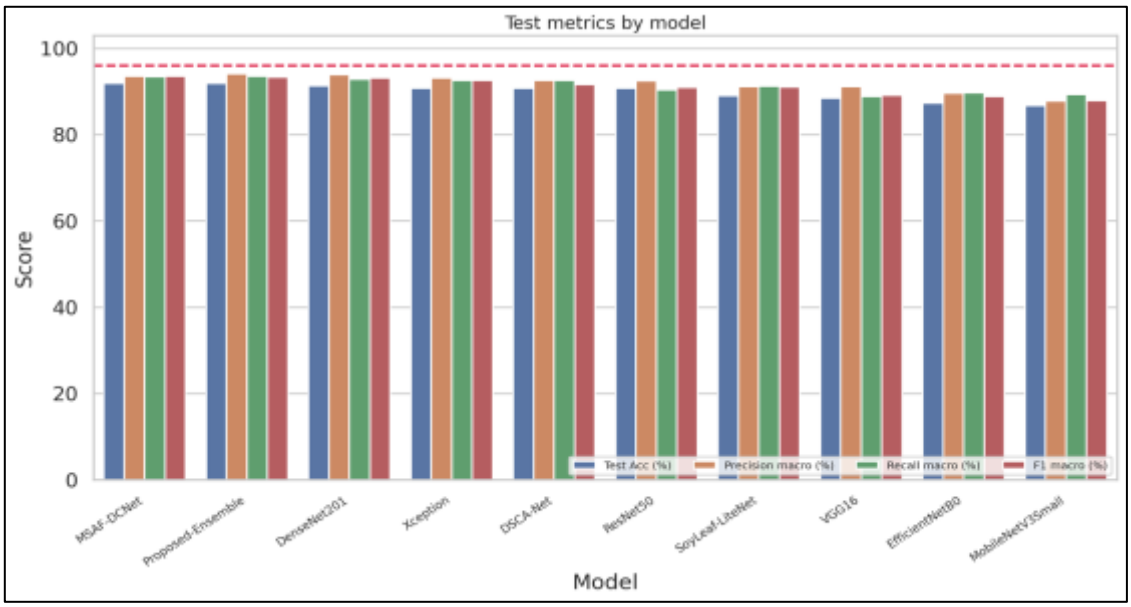



**Fig 13. Test accuracy and macro-averaged precision, recall and F1 by model. Macro scores exceed accuracy for every configuration, indicating stronger performance on the rare tail classes than on the well-populated ones.**

### 5.4 Threshold-Independent Discrimination and Calibration

Because accuracy is a threshold-dependent summary that can conflate ranking quality with decision-boundary placement, Table 6 reports ROC-AUC, PR-AUC and chance-corrected agreement together with log loss. Fig 9 presents per-class ROC and precision–recall curves for the four proposed configurations and Fig 10 the cross-model macro-average ROC comparison.

**Table 6. Threshold-independent, agreement and calibration metrics**

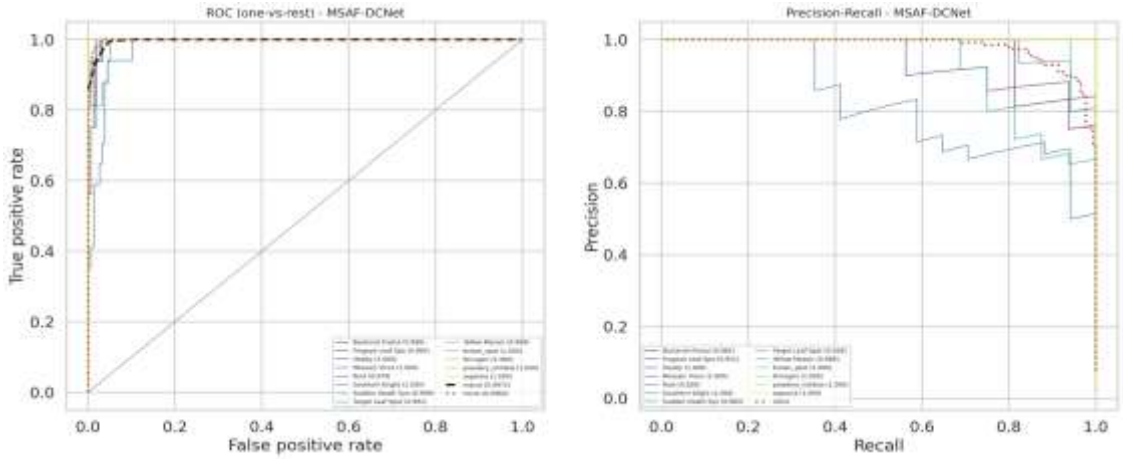
| Model | ROC-AUC (macro) | ROC-AUC (micro) | PR-AUC (macro) | MCC | Cohen's κ | Log Loss |
|---|---|---|---|---|---|---|
| **MSAF-DCNet** | **0.9966** | **0.9982** | **0.9723** | **0.9117** | **0.9116** | **0.2287** |
| **Proposed-Ensemble** | **0.9976** | **0.9987** | **0.9797** | **0.9129** | **0.9117** | **0.2293** |
| DenseNet201 | 0.9966 | 0.9983 | 0.9697 | 0.9064 | 0.9053 | 0.2074 |
| Xception | 0.9963 | 0.9980 | 0.9725 | 0.8998 | 0.8990 | 0.2557 |
| **DSCA-Net** | **0.9967** | **0.9974** | **0.9740** | **0.9015** | **0.8991** | **0.2821** |
| ResNet50 | 0.9965 | 0.9976 | 0.9694 | 0.8995 | 0.8990 | 0.2382 |
| **SoyLeaf-LiteNet** | **0.9947** | **0.9953** | **0.9699** | **0.8805** | **0.8801** | **0.3057** |
| VGG16 | 0.9959 | 0.9967 | 0.9672 | 0.8755 | 0.8739 | 0.2876 |
| EfficientNetB0 | 0.9945 | 0.9961 | 0.9609 | 0.8637 | 0.8612 | 0.3178 |
| MobileNetV3Small | 0.9928 | 0.9949 | 0.9545 | 0.8562 | 0.8550 | 0.3668 |

*Lower log loss indicates better probabilistic calibration; all other columns are higher-is-better.*

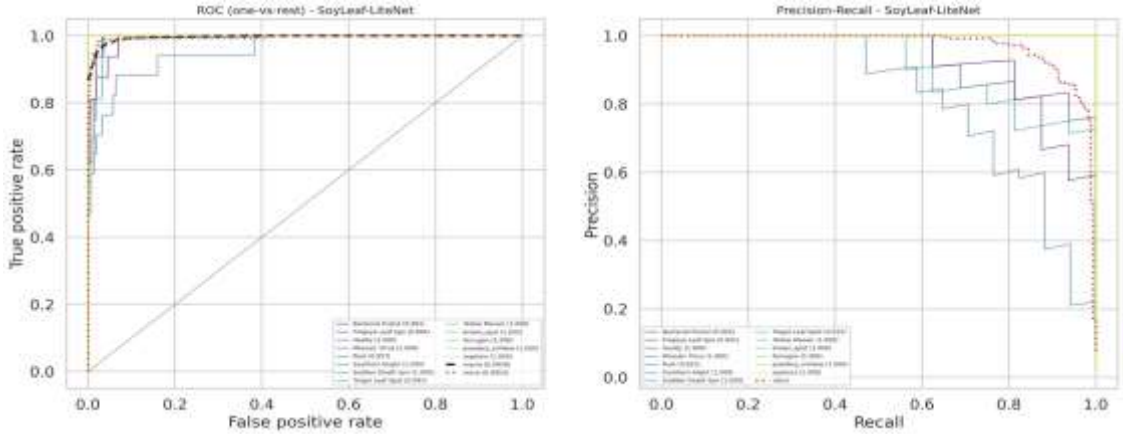
All ten configurations achieve macro ROC-AUC above 0.99, with the Proposed-Ensemble highest at 0.9976 and MobileNetV3Small lowest at 0.9928. The compression of this entire range into 0.0048 AUC units, against a 5.17-point spread in accuracy over the same models, carries an important implication: the evaluated networks differ far more in decision-threshold placement and calibration than in their underlying capacity to rank positives above negatives. A substantial fraction of the accuracy gap separating the proposed models from the weaker baselines is therefore recoverable through threshold tuning or class-prior correction, and part of the proposed architectures' advantage consists in producing better-calibrated posteriors without such post-hoc adjustment.

Macro PR-AUC, which is the more discriminating criterion under imbalance, separates the configurations more informatively than ROC-AUC does. The Proposed-Ensemble leads at 0.9797, followed by DSCA-Net (0.9740), Xception (0.9725) and MSAF-DCNet (0.9723); the 0.0074 margin between the ensemble and MSAF-DCNet is fifteen times the corresponding ROC-AUC margin. This is where the ensemble earns its computational cost: per-class inspection of Fig 9 shows the gain concentrated precisely on the hardest categories, with *Rust* average precision rising from 0.828 for MSAF-DCNet and 0.875 for DSCA-Net to 0.899 for the ensemble, and *Target Leaf Spot* from 0.926 and 0.909 respectively to 0.941. The constituent models therefore make partially independent errors on the difficult cluster, and soft-voting exploits that independence though, as Section 5.6 shows, the residual ambiguity is too strong for fusion to resolve it entirely.

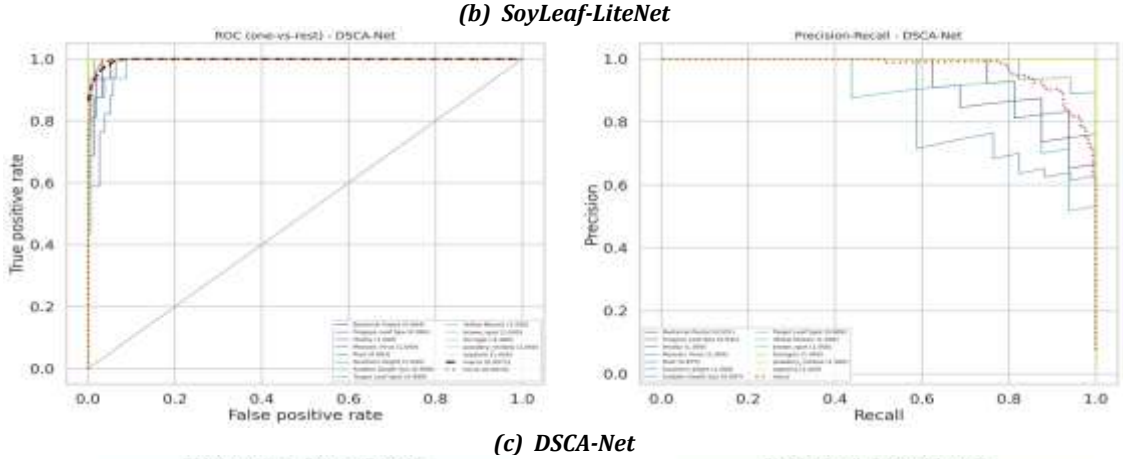
MCC and Cohen's κ, which correct for chance agreement under skewed priors, preserve the ranking established by macro-F1 and confirm that no configuration achieves its accuracy through majority-class bias: the ensemble attains the highest MCC at 0.9129 and MSAF-DCNet the highest κ at 0.9116. One result runs counter to the general ordering and is reported here for completeness. DenseNet201 attains the lowest log loss of any model at 0.2074, ahead of MSAF-DCNet (0.2287) and the ensemble (0.2293), despite ranking third on accuracy. DenseNet201 is thus the best-calibrated model in the comparison even though it is not the most accurate one; where downstream decisions consume predicted probabilities rather than hard labels for example in risk-weighted spray-scheduling this property may outweigh a 0.57-point accuracy deficit. Among the proposed architectures, DSCA-Net's comparatively high log loss of 0.2821 reflects the over-confident misassignment of *Rust* evidence analysed in Section 5.6.



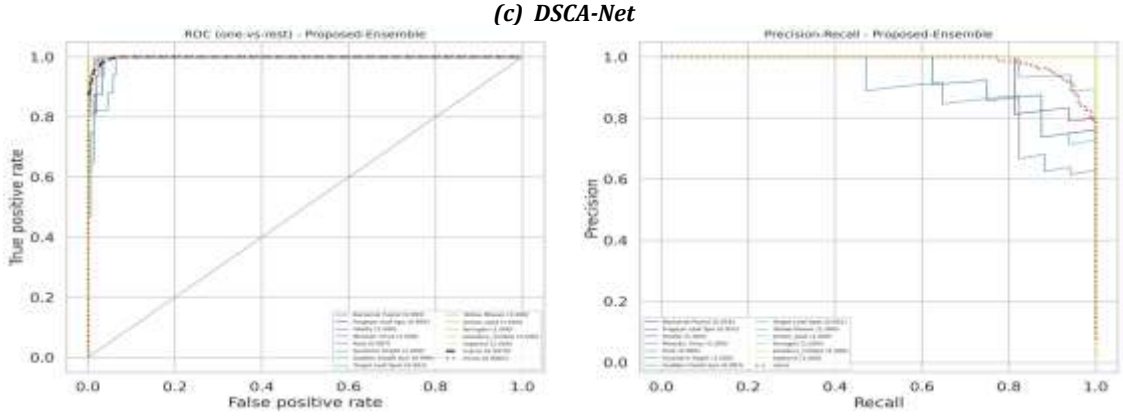

*(a) MSAF-DCNet*




*(b) SoyLeaf-LiteNet*




*(c) DSCA-Net*




*(d) Proposed-Ensemble*

**Fig 9. One-vs-rest ROC and precision–recall curves for the four proposed configurations. Rust is the only class falling below 0.98 ROC-AUC in any panel, and its average precision (0.823–0.899) is far beneath the near-unity values attained elsewhere — the precision–recall view separates the configurations roughly fifteen times more sharply than ROC-AUC does.**

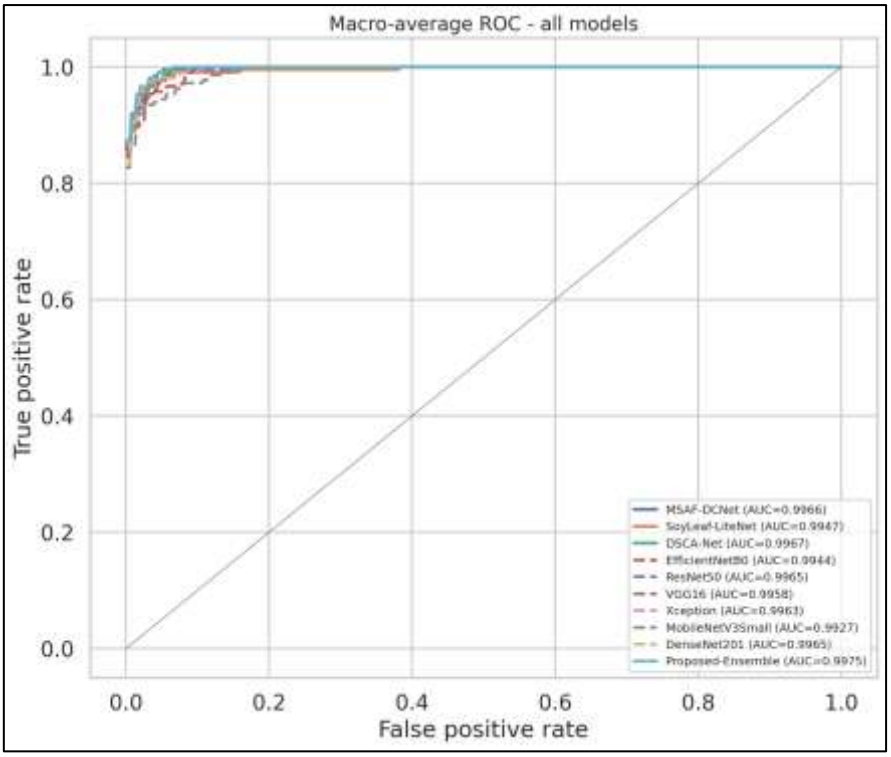



**Fig 10. Macro-average ROC comparison across all ten configurations. The entire range spans 0.0048 AUC units against a 5.17-point spread in accuracy, indicating that the models differ far more in threshold placement and calibration than in underlying ranking capacity.**

### 5.5 Class-Wise Performance

Table 7 compares per-class F1 across the four proposed configurations and Table 8 reports the complete per-class report for MSAF-DCNet, the best-performing single model. Fig 8 presents the corresponding count-based and row-normalised confusion matrices.

**Table 7. Per-class F1-score for the four proposed configurations**

| Class | Support | MSAF-DCNet | SoyLeaf-LiteNet | DSCA-Net | Ensemble |
|---|---|---|---|---|---|
| **Bacterial Pustule** | **16** | **0.8125** | **0.7879** | **0.8750** | **0.8387** |
| Frogeye Leaf Spot | 16 | 0.9091 | 0.8387 | 0.8750 | 0.8750 |
| Healty | 17 | 1.0000 | 0.9714 | 1.0000 | 0.9714 |
| Mossaic Virus | 3 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **Rust** | **17** | **0.7429** | **0.6667** | **0.6923** | **0.7586** |
| Southern blight | 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Sudden Death Syndrome | 17 | 0.9091 | 0.9375 | 0.9091 | 0.9091 |
| **Target Leaf Spot** | **16** | **0.8125** | **0.7429** | **0.7317** | **0.8205** |
| Yellow Mosaic | 17 | 0.9697 | 0.9697 | 0.9697 | 0.9697 |
| brown_spot | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| ferrugen | 10 | 1.0000 | 0.9524 | 1.0000 | 1.0000 |
| powdery_mildew | 21 | 1.0000 | 0.9756 | 1.0000 | 1.0000 |
| septoria | 3 | 1.0000 | 1.0000 | 0.8571 | 1.0000 |
| **Macro average** | **174** | **0.9351** | **0.9110** | **0.9161** | **0.9341** |

*Shaded rows denote the three mutually confusable classes analysed in Section 5.6. All remaining classes exceed F1 = 0.90 for every configuration.*

The error structure is strikingly non-uniform. Of the thirteen categories, **seven are recovered at or near ceiling by every proposed configuration**: *Healty*, *Mossaic Virus*, *Southern blight*, *brown_spot*, *ferrugen*, *powdery_mildew* and *septoria* attain F1 = 1.000 under MSAF-DCNet, with *Yellow Mosaic* at 0.970 and *Sudden Death Syndrome* and *Frogeye Leaf Spot* at 0.909. Only three classes fall below F1 = 0.85 for any configuration, and those three account for the overwhelming majority of all misclassifications.

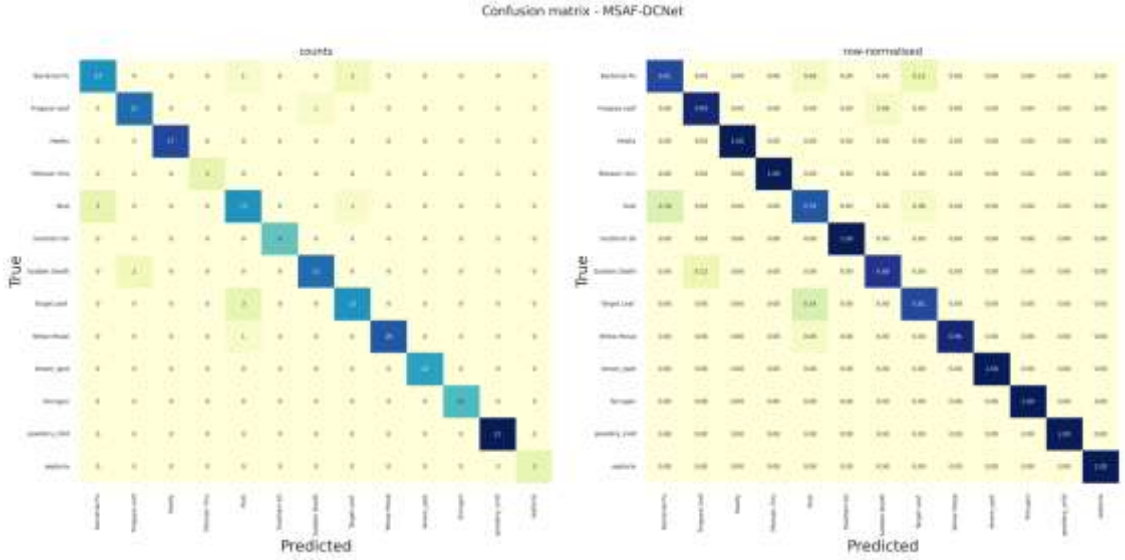
It must be acknowledged that the perfectly separated classes correspond substantially to the background-suppressed and colour-checker acquisition regimes identified in Section 4.1, whereas every substantive error occurs among the unconstrained in-field classes. Two explanations are compatible with this pattern:

the models may be exploiting background cues for the former group, or the former group may simply present more distinctive and less mutually similar lesion morphology. The saliency evidence in Section 5.7 favours the second explanation for the proposed architectures, but the confound cannot be fully excluded from this corpus alone and is identified in Section 5.9 as a limitation requiring targeted ablation.

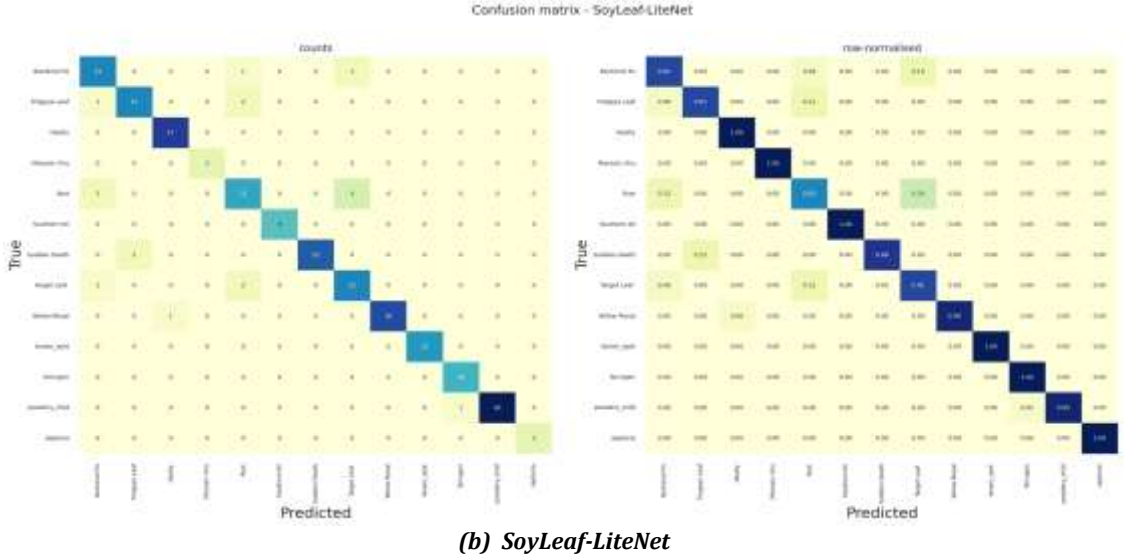
**Table 8. Complete per-class classification report for MSAF-DCNet**

| Class | Support | Precision | Recall | F1-Score |
|---|---|---|---|---|
| **Bacterial Pustule** | **16** | **0.8125** | **0.8125** | **0.8125** |
| Frogeye Leaf Spot | 16 | 0.8824 | 0.9375 | 0.9091 |
| Healty | 17 | 1.0000 | 1.0000 | 1.0000 |
| Mossaic Virus | 3 | 1.0000 | 1.0000 | 1.0000 |
| **Rust** | **17** | **0.7222** | **0.7647** | **0.7429** |
| Southern blight | 9 | 1.0000 | 1.0000 | 1.0000 |
| Sudden Death Syndrome | 17 | 0.9375 | 0.8824 | 0.9091 |
| **Target Leaf Spot** | **16** | **0.8125** | **0.8125** | **0.8125** |
| Yellow Mosaic | 17 | 1.0000 | 0.9412 | 0.9697 |
| brown_spot | 12 | 1.0000 | 1.0000 | 1.0000 |
| ferrugen | 10 | 1.0000 | 1.0000 | 1.0000 |
| powdery_mildew | 21 | 1.0000 | 1.0000 | 1.0000 |
| septoria | 3 | 1.0000 | 1.0000 | 1.0000 |
| **Macro average** | **174** | **0.9359** | **0.9347** | **0.9351** |
| **Weighted average** | **174** | **0.9215** | **0.9195** | **0.9202** |

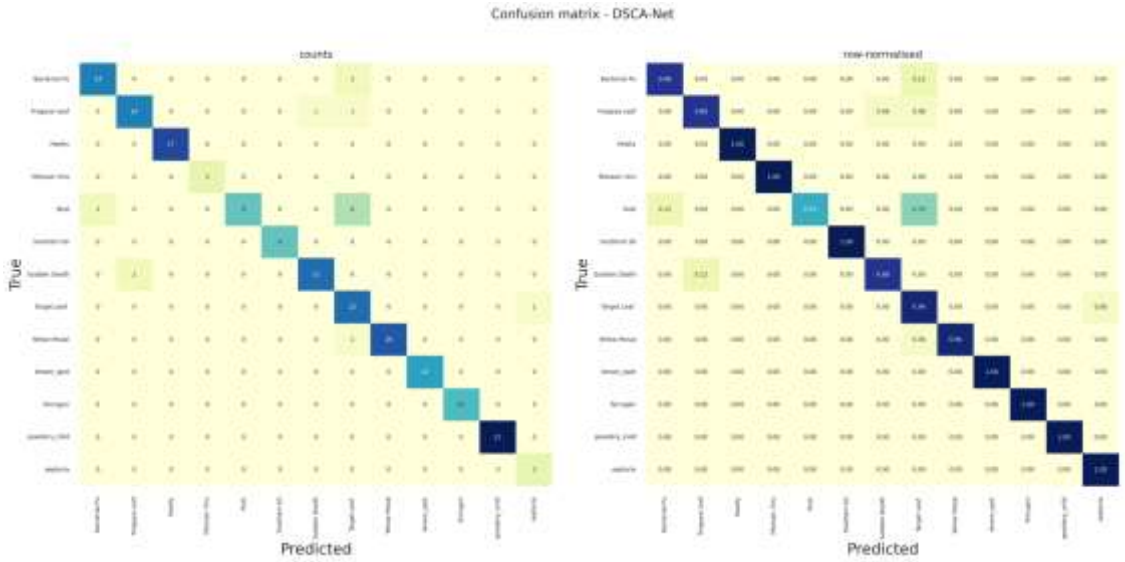
*Overall test accuracy = 0.9195 (160 of 174 images correctly classified).*



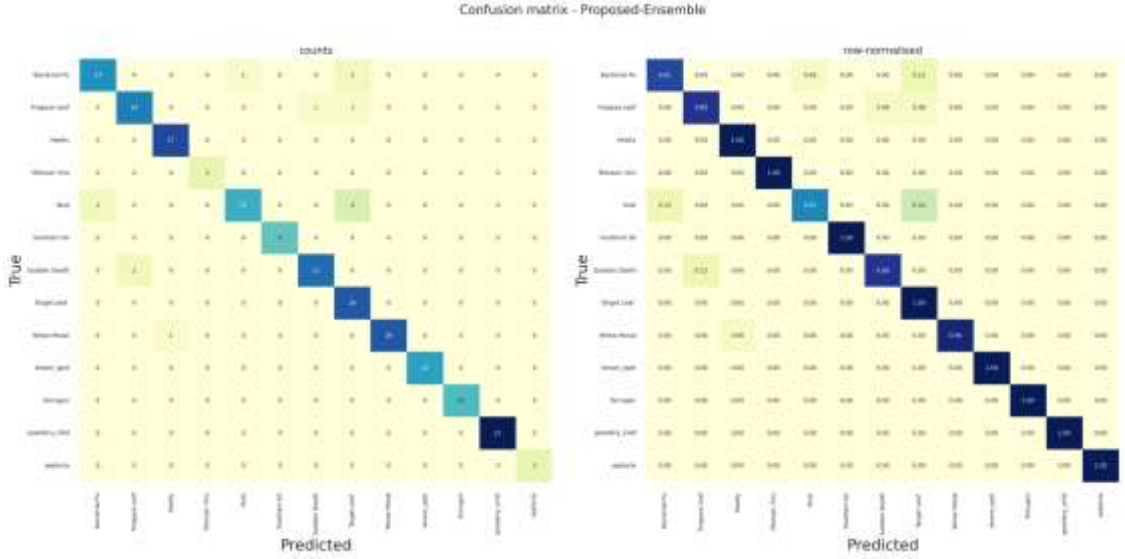

***(a) MSAF-DCNet***




***(b) SoyLeaf-LiteNet***




*(c) DSCA-Net*




*(d) Proposed-Ensemble*

**Fig 8. Count-based and row-normalised confusion matrices on the 174-image test partition. The error structure is markedly non-uniform: seven categories are recovered at ceiling by every configuration, while rust, target leaf spot and bacterial pustule form a mutually confusable cluster that accounts for the majority of all misclassifications.**

### 5.6 Anatomy of the Dominant Error Mode

The residual error in this benchmark is not distributed noise but a single, phytopathologically coherent confusion cluster linking *Rust*, *Target Leaf Spot* and *Bacterial Pustule*. Table 9 decomposes precision and recall for these three classes across the four proposed configurations, and this decomposition proves more diagnostic than the aggregate F1 values of Table 7.

**Table 9. Precision–recall decomposition for the three mutually confusable classes**

| Class | Metric | MSAF-DCNet | SoyLeaf-LiteNet | DSCA-Net | Ensemble |
|---|---|---|---|---|---|
| Rust | Precision | 0.7222 | 0.6875 | 1.0000 | 0.9167 |
| Rust | Recall | 0.7647 | 0.6471 | 0.5294 | 0.6471 |
| Target Leaf Spot | Precision | 0.8125 | 0.6842 | 0.6000 | 0.6957 |
| Target Leaf Spot | Recall | 0.8125 | 0.8125 | 0.9375 | 1.0000 |
| Bacterial Pustule | Precision | 0.8125 | 0.7647 | 0.8750 | 0.8667 |
| Bacterial Pustule | Recall | 0.8125 | 0.8125 | 0.8750 | 0.8125 |

*Read row-pairs together: a high-precision / low-recall pair indicates a class that is under-predicted, its evidence absorbed by a neighbouring category.*

*Rust* is the principal bottleneck across every architecture, and Table 9 reveals that the failure takes a qualitatively different form in different models. **DSCA-Net attains perfect precision on* Rust *(1.0000) at a*

*recall of only 0.5294: it never assigns the* Rust *label incorrectly, but it recovers just 9 of 17 true* Rust *images. The eight lost images are absorbed almost entirely by* Target Leaf Spot*, whose precision collapses correspondingly to 0.6000 against a recall of 0.9375. This is not a discrimination failure in the ranking sense DSCA-Net's* Rust *one-vs-rest AUC is 0.983 but a decision-boundary placement failure, in which the* Target Leaf Spot *logit systematically dominates on ambiguous evidence. The implication is practical: DSCA-Net's* Rust* performance is substantially recoverable by class-prior recalibration without retraining, a conclusion that aggregate accuracy alone would not have surfaced.

MSAF-DCNet distributes the same underlying ambiguity more symmetrically, with *Rust* precision 0.7222 against recall 0.7647 and *Target Leaf Spot* balanced at 0.8125 on both metrics. It is this balance, rather than any superior separation of the confused pair, that produces its macro-F1 advantage. The ensemble inherits DSCA-Net's bias in attenuated form: it drives *Target Leaf Spot* recall to a perfect 1.0000 while *Target Leaf Spot* precision remains at 0.6957, and lifts *Rust* precision to 0.9167 at a recall of 0.6471. The ensemble's headline macro-precision advantage is therefore not uniform across classes it is achieved despite, not because of, its behaviour on *Target Leaf Spot* and this qualification should accompany any claim made for the fusion strategy.

The confusion is phytopathologically well-founded. Soybean rust (*Phakopsora pachyrhizi*), target leaf spot (*Corynespora cassiicola*) and bacterial pustule (*Xanthomonas axonopodis* pv. *glycines*) all present as small, discrete, reddish-brown to tan foliar lesions on an otherwise green lamina during early to mid infection. The features that reliably separate them the raised uredinial pustules of rust, the concentric ring pattern of target spot, and the pale raised centre of the bacterial pustule subtend only a few pixels at the 320 px working resolution and are further degraded by the variable illumination and focal distance of field acquisition. The confusion is also bidirectional rather than a one-sided prior bias: MSAF-DCNet assigns 3 of 16 *Target Leaf Spot* images and 1 of 17 *Yellow Mosaic* images to *Rust*, confirming a genuinely entangled decision region.

A second and considerably weaker confusion links *Sudden Death Syndrome* to *Frogeye Leaf Spot*, with 2 of 17 SDS images misassigned by all four configurations a stable recall of 0.8824 across architecturally distinct models. Both conditions exhibit interveinal chlorosis with necrotic centres in field imagery. The invariance of this error across models again indicates intrinsic visual ambiguity rather than a model-specific artefact. Beyond these two clusters, the only isolated errors are SoyLeaf-LiteNet's single *powdery_mildew* image assigned to *ferrugen* and DSCA-Net's single *Target Leaf Spot* image assigned to *septoria*, the latter reducing *septoria* precision to 0.750 purely through one false positive against a support of three.

**5.7 Interpretability Analysis**

**Fig 14** presents gradient-saliency overlays for MSAF-DCNet, DSCA-Net and SoyLeaf-LiteNet computed on a common set of eight test images, providing a direct test of the background-shortcut hypothesis raised in Section 4.1.

For MSAF-DCNet the salient energy is sparse and lesion-localised. On the *Bacterial Pustule* sample the strongest gradients coincide with the discrete necrotic spots on the lamina rather than with the surrounding canopy, and on the *powdery_mildew* sample where the leaf is segmented onto a null black field activation is confined to the leaf boundary and surface texture, with essentially zero response over the background. This is the desired behaviour and indicates that the multi-scale attention fusion mechanism selects lesion-scale evidence rather than acquisition-regime cues. DSCA-Net produces a denser, more diffuse saliency field distributed across the whole lamina, consistent with its dual-stream channel attention integrating contextual information more broadly. SoyLeaf-LiteNet yields the most scattered maps, with visible response along the image border of the *brown_spot* sample, suggesting that its reduced capacity leaves it more susceptible to background correlation. This ordering of interpretability quality mirrors the accuracy ordering of the three proposed models and offers a mechanistic account of the compact model's residual gap.


***(a) MSAF-DCNet***


*(b) DSCA-Net*


*(c) SoyLeaf-LiteNet*

**Fig 14. Gradient-saliency overlays computed on a common set of test images. MSAF-DCNet concentrates salient energy on lesion structures and shows essentially no response over the null background of the segmented samples; DSCA-Net produces a more diffuse field; SoyLeaf-LiteNet shows the most scattered attribution, including response along the image border. This ordering mirrors the accuracy ordering of the three models.**

### 5.8 Qualitative Prediction Analysis and Confidence Reliability

**Fig 15** shows actual-versus-predicted results for MSAF-DCNet over 24 randomly drawn test images, with correct predictions bordered in green and errors in red. Twenty-one of the twenty-four samples are correctly classified, and critically predicted confidence is strongly diagnostic of correctness.

Correct predictions are issued at high confidence, spanning 0.86 to 1.00 with fourteen samples above 0.95. All three errors, by contrast, occur at low confidence: *Rust* misread as *Bacterial Pustule* at 0.45, *Yellow Mosaic* misread as *Rust* at 0.62, and *Bacterial Pustule* misread as *Target Leaf Spot* at 0.61. Every error falls within the confusion cluster identified in Section 5.6, and every error is flagged by a posterior below 0.65 while no correct prediction falls below 0.86. This separation is operationally valuable: the selective-prediction rule of Eq. (24), at $\tau = 0.70$, would refer the uncertain cases for human agronomist review while retaining the confidently correct majority, converting the model's principal weakness from silent misdiagnosis into a bounded and manageable referral rate. Read together with the 100.00% top-3 accuracy of the ensemble reported in Table 5, this suggests that a deployment combining an abstention threshold with a ranked candidate shortlist would substantially exceed the practical utility implied by the 91.95% top-1 figure.

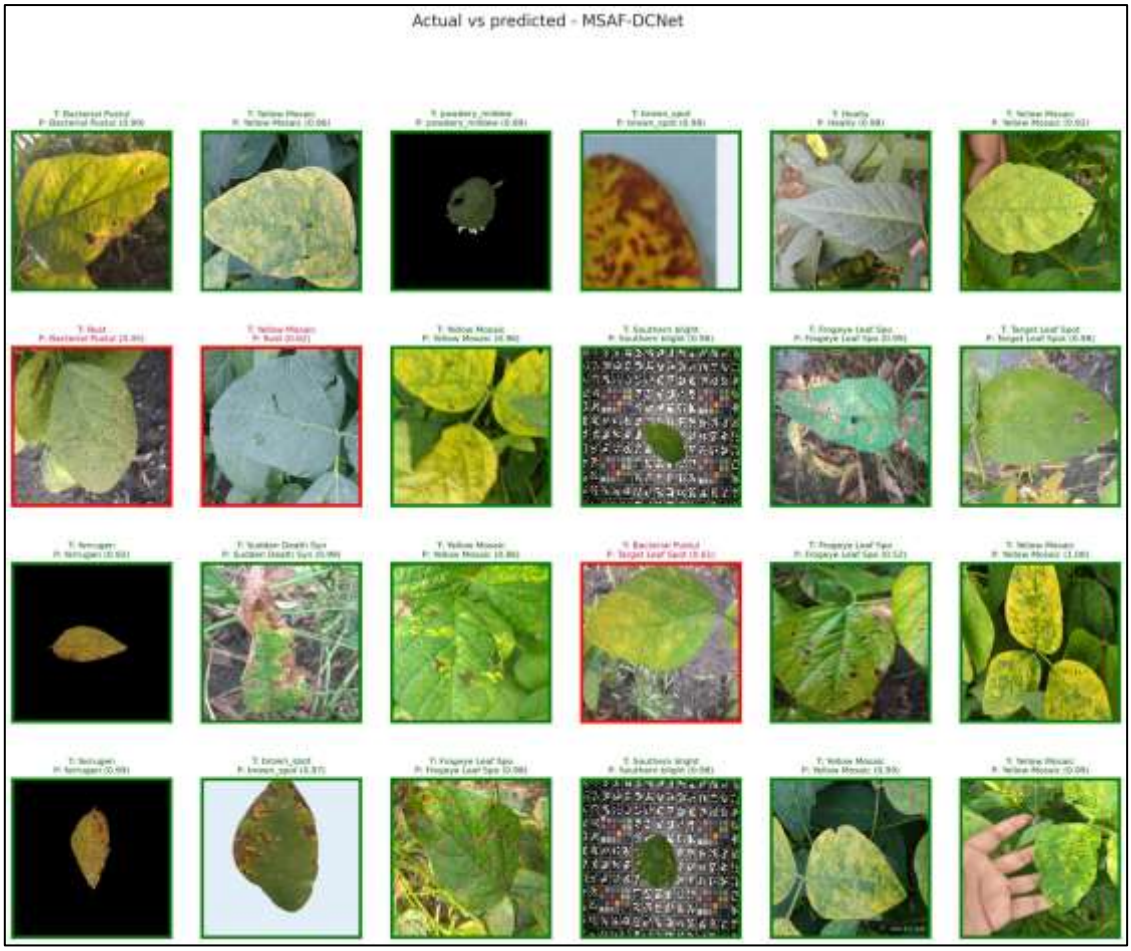



**Fig 15. Actual versus predicted results for MSAF-DCNet over 24 randomly drawn test images; green borders denote correct predictions and red borders errors. Correct predictions span 0.86–1.00 confidence while all three errors fall below 0.65, so a threshold at 0.70 separates them cleanly — the empirical basis for the abstention rule of Eq. (24).**

### 5.9 Statistical Considerations and Limitations

Four qualifications constrain the strength of the claims that this evidence supports, and are stated here rather than deferred.

**Test-set size and effect magnitude.** The test partition contains 174 images, so a single image is worth 0.575 percentage points of accuracy. The Wilson 95% confidence interval for MSAF-DCNet's 91.95% is [86.9%, 95.1%], and that for MobileNetV3Small's 86.78% is [80.9%, 91.0%] intervals that overlap substantially. The 0.57-point margin separating MSAF-DCNet from DenseNet201 corresponds to exactly one test image and is not statistically distinguishable from chance at this sample size; the same applies to the three-way tie at 90.80% among Xception, DSCA-Net and ResNet50. The defensible claims are therefore that the proposed architectures are *competitive with or superior to* established backbones and that the accuracy ordering is *consistent with* the macro-F1, MCC and PR-AUC orderings not that any individual pairwise ranking is significant. A McNemar test on paired predictions, together with bootstrap confidence intervals over the test partition, is recommended before the final ranking claim is made.

**Single-seed evaluation.** All results reported here derive from a single training run per architecture with a fixed data partition; the seed-variance column of the experimental log is empty. Reported differences of under approximately one percentage point should therefore be regarded as within run-to-run variance. Repeating each configuration across three to five seeds and reporting mean ± standard deviation, or adopting stratified k-fold cross-validation, would materially strengthen the comparison and is the single most valuable addition available at modest computational cost.

**Acquisition confound.** As established in Sections 4.1 and 5.5, acquisition regime correlates with class identity in this corpus, and the perfectly separated classes are disproportionately those captured against uniform or calibrated backgrounds. The saliency evidence of Section 5.7 indicates that MSAF-DCNet attends to lesion morphology rather than background, but saliency is suggestive rather than conclusive. A background-ablation experiment retraining on segmented or background-randomised images and measuring the accuracy differential would settle the question directly and is recommended prior to any claim of field generalisability.

**Unmet design target and its remedy.** No configuration reaches the 96% target marked in Figs 7 and 13, and Section 5.6 attributes the shortfall almost entirely to the rust/target-spot/pustule cluster. Since the *Rust* one-vs-rest AUC remains at 0.979–0.987 while its recall falls as low as 0.529, the deficiency is one of resolution and decision-boundary placement rather than of representational capacity. Three targeted interventions follow directly from this diagnosis: increasing the working resolution or introducing lesion-region cropping so that pustule and ring texture is resolvable; applying class-prior correction or cost-sensitive thresholding to the *Rust*/*Target Leaf Spot* pair, which requires no change to Eq. (19), which Table 9 indicates would recover a substantial fraction of DSCA-Net's lost recall without retraining; and targeted augmentation or additional acquisition for the *Rust* class specifically. These are pursued in Section 6.

## 6. Conclusion

This article has developed and evaluated a three-architecture framework for soybean foliar disease classification, motivated by the observation that the discriminations which matter most in this task occur at a spatial scale that conventional transfer-learning pipelines discard. MSAF-DCNet addresses that mismatch directly, taking three resolution taps at strides of 8, 16 and 32 pixels, recalibrating each with convolutional block attention and fusing concatenated average and maximum statistics. On a 13-class, 1,158-image benchmark it attains 91.95% test accuracy, 93.51% macro-F1 and a macro ROC-AUC of 0.9966, outperforming six established ImageNet-pretrained backbones trained under an identical protocol. Two results carry more practical weight than the headline ranking. SoyLeaf-LiteNet recovers 89.08% accuracy 96.9% of the flagship model's performance using 7.27 M parameters and one third of the training time, and outranks three standard baselines in the process; for field hardware, that position on the cost–accuracy curve matters more than the 2.87-point deficit. And the confidence-weighted ensemble, while it reproduces MSAF-DCNet's accuracy exactly rather than improving on it, attains the best macro-precision (94.12%), the best macro PR-AUC (0.9797) and 100% top-3 accuracy, which is the more relevant criterion for a system that presents a ranked shortlist to a human agronomist.

The most useful finding, however, is diagnostic rather than competitive. Almost the entire residual error resolves into one phytopathologically coherent cluster linking rust, target leaf spot and bacterial pustule, and the per-class precision–recall decomposition shows this to be a decision-threshold failure rather than a representational one: DSCA-Net achieves perfect precision on rust while recovering only 52.94% of it, with a one-vs-rest AUC of 0.983. A model that ranks rust correctly but refuses to predict it is recoverable by class-prior correction without retraining, and this distinction invisible in aggregate accuracy is the practical route to closing the gap to the 96% design target that no configuration in this study reached.

Finally, the evaluation is reported with its limits stated. On a 174-image test partition a single image is worth 0.575 percentage points, the Wilson 95% interval for the leading model spans [86.9%, 95.1%], and the margin separating it from the strongest baseline corresponds to exactly one image. The defensible claim is that the proposed architectures are competitive with or superior to established backbones and that the ordering is consistent across accuracy, macro-F1, MCC and PR-AUC not that any individual pairwise ranking is statistically significant.

## References


1. Y. Zhang, R. Bao, M. Guan, Z. Wang, L. Wang, X. Cui, and Y. Wang, “A lightweight deep convolutional neural network development for soybean leaf disease recognition,” Frontiers in Plant Science, vol. 16, art. 1655564, 2025.
2. A. K. Shinwari, S. Hameed, M. S. Akhter, F. U. Hassan, and A. Sani, “A vision for the future: The next wave of innovation in precision agriculture,” in Advancing Precision Agriculture With AI and IoT. IGI Global Scientific Publishing, 2026, pp. 427–456.
3. V. Pandit, M. Sravya, C. Ravali, S. Fiaz, A. Tariq, and S. Chatterjee, “Precision agriculture in the era of technological advancement,” in Artificial Intelligence and Data Sciences for Precision Agriculture. Cham, Switzerland: Springer Nature Switzerland, 2026, pp. 159–186.
4. B. Gowshika, “Satellite based agriculture field monitoring and data analysis system,” in Proc. 6th Int. Conf. Trends Mater. Sci. Inventive Mater. (ICTMIM), 2026, pp. 33–38.
5. A. Mulakaledu, B. Swathi, M. M. Jadhav, S. M. Shukri, V. Bakka, and P. Jangir, “Satellite image–based ecosystem monitoring with sustainable agriculture analysis using machine learning model,” Remote Sensing in Earth Systems Sciences, pp. 1–10, 2024.
6. S. Choudhury, “Advanced technology of using satellite data fuels in agriculture,” Chronicle of Bioresource Management, vol. 8, no. 1, pp. 029–034, 2024.
7. Chatrabhuj and K. Meshram, “Incremental learning model for sustainable agricultural land assessment using multimodal satellite data,” Int. J. Remote Sens., vol. 45, no. 22, pp. 8622–8648, 2024.
8. R. Castro, “Prediction of yield and diseases in crops using vegetation indices through satellite image processing,” in Proc. IEEE Technol. Eng. Manage. Soc. (TEMSCON LATAM), 2024, pp. 1–6.
9. S. Gore, D. Patil, N. Mahankale, and S. Gore, “Satellite imaging for precision agriculture enhancing crop management, soil condition and yield prediction,” in Emerging Trends in Smart Societies. Routledge, 2024, pp. 434–437.
10. J.-K. Yu, J. Ahn, G. D. Han, H.-M. Kang, H. Jo, and Y. S. Chung, “Utilization of satellite technologies for agriculture,” J. Environ. Sci. Int., vol. 33, no. 7, pp. 547–552, 2024.

11. O. Opryshko, N. Pasichnyk, N. Kiktev, A. Dudnyk, T. Hutsol, K. Mudryk, P. Herbut, P. Łyszczarz, and V. Kukharets, "European Green Deal: Satellite monitoring in the implementation of the concept of agricultural development in an urbanized environment," Sustainability, vol. 16, no. 7, art. 2649, 2024.
12. N. Efremova, J. C. Foley, A. Unagaev, and R. Karimi, "AI for sustainable agriculture and rangeland monitoring," in The Ethics of Artificial Intelligence for the Sustainable Development Goals. Cham, Switzerland: Springer, 2023, pp. 399–422.
13. J. Segarra, "Satellite imagery in precision agriculture," in Digital Agriculture: A Solution for Sustainable Food and Nutritional Security. Cham, Switzerland: Springer, 2024, pp. 325–340.
14. D. Wüpper and W. A. Oluoch, "Satellite data in agricultural and environmental economics: Theory and practice," Agricultural and Environmental Economics, preprint, 2024.
15. A. E. Badillo-Márquez, I. Pardo-Escandón, A. A. Aguilar-Lasserre, C. G. Moras-Sánchez, and R. Flores-Asis, "Intelligent system based on a satellite image detection algorithm and a fuzzy model for evaluating sugarcane crop quality by predicting uncertain climatic parameters," J. Agricultural Engineering, vol. 55, no. 4, 2024.
16. A. A. Kzar, "Change detection using multispectral satellite images for sustainable development in regions of Al-Najaf Al-Ashraf, Iraq," J. Kufa-Physics, vol. 16, no. 1, pp. 42–55, 2024.
17. B. Victor, A. Nibali, and Z. He, "A systematic review of the use of deep learning in satellite imagery for agriculture," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 2024.
18. S. Nair, S. Sharifzadeh, and V. Palade, "Farmland segmentation in Landsat 8 satellite images using deep learning and conditional generative adversarial networks," Remote Sensing, vol. 16, no. 5, art. 823, 2024.
19. K. Anderson, "Detecting environmental stress in agriculture using satellite imagery and spectral indices," Environmental Monitoring and Assessment, preprint, 2024.
20. A. Balasundaram, A. B. Abdul Aziz, A. Gupta, A. Shaik, and M. S. Kavitha, "A fusion approach using GIS, green area detection, weather API and GPT for satellite image-based fertile land discovery and crop suitability," Scientific Reports, vol. 14, no. 1, art. 16241, 2024.
21. P. Lemenkova, "Deep learning methods of satellite image processing for monitoring of flood dynamics in the Ganges Delta, Bangladesh," Water, vol. 16, art. 1141, 2024.
22. S. Trivedi, P. V. Vinod, B. Chandrasekaran, M. K. Nagashree, S. R. Subramoniam, V. B. Manjula, and A. Singh, "Geospatial assessment of agroforestry land use systems using very high-resolution satellite images and artificial intelligence," Agroforestry Systems, vol. 98, no. 8, pp. 2875–2895, 2024.
23. M. A. Haq, "Intelligent sustainable agricultural water practice using multi-sensor spatiotemporal evolution," Environmental Technology, vol. 45, no. 12, pp. 2285–2298, 2024.
24. A. Raihan, "A systematic review of geographic information systems (GIS) in agriculture for evidence-based decision making and sustainability," Global Sustainability Research, vol. 3, no. 1, pp. 1–24, 2024.
25. S. S. Rana, I. I. Rana, M. J. Khan, S. Habib, R. M. Saleem, H. M. Haroon, and H. Irfan, "Predicting and identifying land features utilising remote sensing satellite imagery and machine learning techniques," Kashf J. Multidisciplinary Research, vol. 1, no. 12, pp. 17–35, 2024.
26. P. Chinnasamy, D. Tejaswini, R. K. Ayyasamy, S. Dhanasekaran, B. S. Kumar, and L. Chandran, "Crop optimization and disease detection using satellite imagery and artificial intelligence," in Proc. 2nd Int. Conf. Intell. Cyber Phys. Syst. Internet Things (ICoICI), 2024, pp. 1531–1535.
27. Chatrabhuj and K. Meshram, "Design of an efficient satellite image analysis model for identification of sustainable construction areas via ground subsidence monitoring and infrastructure monitoring operations," Indian Geotechnical Journal, preprint, pp. 1–16, 2024.
28. S. Getahun, H. Kefale, and Y. Gelaye, "Application of precision agriculture technologies for sustainable crop production and environmental sustainability: A systematic review," The Scientific World Journal, vol. 2024, no. 1, art. 2126734, 2024.
29. B. L. Mahesh and S. U. Shenoy, "Advances in crop classification using satellite imagery: A comprehensive review," in Proc. IEEE Int. Conf. Comput. Vis. Mach. Intell. (CVMI), 2024, pp. 1–7.
30. F. L. P. de Souza, L. S. Shiratsuchi, M. A. Dias, M. R. Barbosa Júnior, T. D. Setiyono, S. Campos, and H. Tao, "A neural network approach employed to classify soybean plants using multi-sensor images," Precision Agriculture, vol. 26, no. 2, art. 32, 2025.
31. Y.-H. Park, S. H. Choi, Y.-J. Kwon, S.-W. Kwon, Y. J. Kang, and T.-H. Jun, "Detection of soybean insect pest and a forecasting platform using deep learning with unmanned ground vehicles," Agronomy, vol. 13, no. 2, art. 477, 2023.
32. E. C. Tetila et al., "Automatic recognition of soybean leaf diseases using UAV images and deep convolutional neural networks," IEEE Geosci. Remote Sens. Lett., vol. 17, no. 5, pp. 903–907, May 2020.
33. B. Wang, Y. Gao, X. Yuan, and S. Xiong, "Local R-symmetry co-occurrence: Characterising leaf image patterns for identifying cultivars," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 19, no. 2, pp. 1018–1031, Mar./Apr. 2022.
34. Y. Hang, X. Meng, and Q. Wu, "Application of improved lightweight network and Choquet fuzzy ensemble technology for soybean disease identification," IEEE Access, vol. 12, pp. 25146–25163, 2024.
35. W. Liu, G. Wu, H. Wang, and F. Ren, "Lightweight pyramidal multi-scale attention residual network for plant disease recognition," IEEE Access, vol. 13, pp. 61526–61536, 2025.
36. N. Farah, N. Drack, H. Dawel, and R. Buettner, "A deep learning-based approach for the detection of infested soybean leaves," IEEE Access, vol. 11, pp. 99670–99679, 2023.
37. B. H. Kim and S. H. Seo, "Soybean leaf disease identification through smart detection using machine learning-convolutional neural network model," Legume Research, vol. 48, no. 6, pp. 1043–1050, 2026.
38. A. Alnuaim, A. Altheneyan, and A. A. AlZubi, "Early leaf disease detection of soybean plants using convolution neural network algorithm," Legume Research, vol. 48, no. 6, pp. 1025–1034, 2026.
39. W. An, J. Yang, Y. Ren, and H. Ni, "High-precision MSFFANet model for soybean leaf disease recognition," Discover Computing, vol. 29, no. 1, art. 330, 2026.

40. X. Li, W. Bian, B. Yang, Q. Yang, W. Cui, J. Liang, and J. Gu, "Soybean leaf disease recognition based on Sem-ResFormer and multimodal large models," Agronomy, vol. 16, no. 12, art. 1132, 2026.
41. R. Yadav, A. Seth, S. Kolhe, S. Kumar, B. Dupare, and M. Kumbhkar, "AgriNet: A compact deep learning pipeline for segmented soybean leaf disease detection in field environments," Procedia Computer Science, vol. 283, pp. 3049–3062, 2026.
42. F. B. Adugna, E. A. Desta, and A. A. Hailu, "Efficient multi-task CNN framework for joint identification of soybean leaf diseases and pesticide presence with explainable AI," Scientific Reports, 2026.
43. X. Li, W. Bian, B. Yang, Y. Li, S. Wang, N. Qin, and H. Sun, "Soybean leaf disease recognition methods based on hyperparameter transfer and progressive fine-tuning of large models," Agronomy, vol. 16, no. 2, art. 218, 2026.
44. S. Nath, S. K. Biswas, S. Majumdar, K. Kumar, and T. Acharjee, "Smart farming using deep learning: Detection of soybean leaf disease from UAV-captured images," in Proc. Int. Conf. Emerging Syst. Intell. Comput. (ESIC), Feb. 2026, pp. 109–114.
45. S. S. Mohanty, V. Maheshwari, P. K. Aithal, and R. D. Jathanna, "A multi-scale supervised contrastive framework for cross-domain soybean disease classification using leaf and UAV imagery," Scientific Reports, 2026.
46. F. Nan, Q. Liu, Y. Hu, and H. Liu, "Soy-Hybrid: A hybrid architecture for accurate and efficient soybean leaf disease recognition," in Proc. Int. Conf. Intell. Comput., Singapore: Springer Nature Singapore, Jul. 2026, pp. 289–300.
47. S. Nath, S. K. Biswas, K. Kumar, D. Tripathi, and S. Majumdar, "A CNN model for soybean leaf disease prediction treating class imbalance problem using diffusion model: A smart farming application," in Proc. 4th Int. Conf. Secure Cyber Comput. Commun. (ICSCCC), May 2026, pp. 1266–1271.