Information Processing in Agriculture xxx (xxxx) xxx

Contents lists available at ScienceDirect

Information Processing in Agriculture

journal homepage: www.keaipublishing.com/en/journals/information-processing-in-agriculture/

Neuro-symbolic AI for rice disease diagnosis with calibrated attention and rule-aware explanations

Chatter Singh [a], Amar Singh [a], Sahraoui Dhelim [b,*]

[a] School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India
[b] School of Computing, Dublin City University, Dublin, Ireland

ARTICLE INFO

Keywords:
Post-hoc explainability
Rice disease diagnosis
Calibrated confidence
CBAM attention
Temperature scaling
Grad-CAM
RDF/OWL rules
Uncertainty and robustness

ABSTRACT

Accurate and trustworthy disease diagnosis from field imagery requires a framework that balances predictive accuracy with calibrated confidence and auditable reasoning. This work benchmarks a diagnostic system coupling an attention-augmented convolutional network (ResNet-34+CBAM) with post-hoc probability calibration and a rule-aware validator. Agronomic symptom rules are encoded in a lightweight RDF/OWL knowledge graph, enabling a post-hoc check that links model predictions to human-readable explanations for auditability. On a rigorously de-duplicated test split of the public PaddyDoctor corpus, the model achieves 95.13% accuracy (weighted F1: 95.14%) with a median latency of 4.6 ms. We analyze the trade-offs of post-hoc calibration: Temperature Scaling, fit on the calibration split (ECE: 1.65%→1.35%), improves the test-set Brier score (0.0760→0.0758) and NLL (0.1573→0.1566) but results in a slight increase in test-set ECE (0.82%→0.94%). A robustness analysis using common corruptions identifies critical failure modes: while resilient to JPEG compression (86.15% accuracy at severity 5), the model is highly vulnerable to brightness shifts (47.72%) and Gaussian blur (32.13%), highlighting the need for domain-specific augmentations. The resulting system provides a comprehensive baseline for combining strong predictive performance with post-hoc calibration and auditable explanations, supporting transparent triage in practical deployments.

1. Introduction

Rice disease diagnosis in practice is frequently performed from handheld or smartphone images captured in uncontrolled field conditions. Unlike laboratory imagery, these inputs exhibit strong variation in illumination (direct sun, shade, backlighting), motion blur from wind or hand tremor, cluttered backgrounds, and compression artifacts from messaging apps. Decisions based on such images (spraying, isolation, replanting, or escalation to laboratory testing) are time-sensitive and carry direct cost and yield implications.

These deployment realities impose requirements that go beyond top-1 accuracy. Agronomists and extension workers need (i) *reliable probabilities* to support threshold-based triage and escalation, (ii) *auditable rationales* that can be communicated and checked against agronomic knowledge, and (iii) *predictable behavior* under common corruptions and distribution shift encountered in the field. In contrast, many off-the-shelf CNN pipelines remain prone to miscalibration and overconfident errors, and saliency visualizations alone rarely satisfy audit- or policy-oriented explanation standards. Neuro-symbolic and knowledge-guided approaches can help close this gap by coupling pattern-recognizing neural models with explicit symptom rules represented in machine-readable form [1–7].

Problem definition. Given an in-the-wild RGB rice-leaf image x, our goal is to output a disease prediction $\hat{d}$ together with a calibrated confidence score suitable for operational thresholds, and an explanation that can be *audited* against symptom-level knowledge (e.g., lesion color, morphology, and spatial patterns). In addition, because field capture conditions are a dominant source of failure in agricultural vision, we explicitly characterize robustness under controlled corruptions that mirror real deployment variability.

In this work, we study rice disease diagnosis from RGB leaf imagery and benchmark a deployment-oriented pipeline that combines attention-augmented recognition, probabilistic calibration, and *post-hoc* rule-aware verification. We fine-tune a ResNet-34 classifier with CBAM attention to promote focus on symptom-bearing regions, then fit temperature scaling on a held-out calibration split to improve probability quality. Predictions are subsequently audited against an RDF/OWL knowledge graph encoding agronomic priors [6–9]. This validator produces a discrete audit state (Aligned/Conflict/NA), rule-alignment scores, and textual rationales designed for decision support and traceable review [10–15].

* Corresponding author.
E-mail addresses: manhas.cs@gmail.com (C. Singh), amar.23318@lpu.co.in (A. Singh), sahraoui.dhelim@dcu.ie (S. Dhelim).

https://doi.org/10.1016/j.inpa.2026.02.006
Received 22 November 2025; Received in revised form 7 February 2026; Accepted 10 February 2026
Available online 11 February 2026

We evaluate on a *rigorously de-duplicated* split of the public *PaddyDoctor* corpus (16,225 field images; 13 classes including Healthy) [1,16–19]. The resulting model achieves 95.13% test accuracy (weighted F1: 95.14%) with a median latency of ∼4.6 ms per image. Beyond headline accuracy, we quantify calibration and robustness. Temperature scaling improves proper scoring rules on the test set (NLL: 0.1573→0.1566; Brier: 0.0760→0.0758) and reduces ECE on the calibration split (1.65%→1.35%), while slightly increasing test-set ECE (0.82%→0.94%), reflecting a known trade-off in post-hoc calibration under distribution shift [12–15,20]. A corruption-based robustness diagnosis reveals critical field-relevant failure modes: performance is comparatively resilient to JPEG compression, but accuracy collapses to **47.72%** under severe brightness shifts and to **32.13%** under Gaussian blur [21–23].

Contributions.

- **Attention-augmented benchmark under a rigorous split.** We report a strong baseline (95.14% weighted F1) for a CBAM–ResNet-34 classifier on a leakage-resistant and de-duplicated *PaddyDoctor* split.
- **Calibration analysis with transparent trade-offs.** We evaluate post-hoc probability calibration and explicitly quantify the ECE–NLL/Brier trade-off, reporting both calibration-split and test-set behavior under the same protocol [12–15,20].
- **Rule-aware explanations for auditable decision support.** We propose a lightweight RDF/OWL knowledge-graph validator that audits CNN predictions, assigns audit states (Aligned/Conflict/NA), and generates domain-grounded rationales suitable for review and escalation [6–9].
- **Robustness diagnosis of field-relevant failures.** Using a standard corruption benchmark, we identify actionable vulnerabilities to photometric shift and blur to motivate domain-specific augmentation and robustness interventions [17,21–23].

The remainder of the paper is organized as follows. Section 2 reviews related work in plant disease vision, calibration, and rule-aware auditing. Section 3 describes the dataset, model, calibration, and evaluation protocol. Section 4 reports quantitative results and deployment-relevant analyses (calibration, triage, robustness, and efficiency), and Section 5 concludes with limitations and future directions.

2. Related work

Deep models for plant disease recognition. Deep learning is the dominant paradigm for image-based crop pathology, with convolutional neural networks (CNNs) remaining a practical default in many field settings due to their accuracy–efficiency trade-off [24,25]. Across rice and broader crops, CNN backbones with lightweight classification heads have repeatedly demonstrated strong symptom recognition under cluttered backgrounds and in-the-wild capture conditions [2,3,17,26–30]. *Limitations in practice:* many studies emphasize accuracy under a single split and do not explicitly control for near-duplicate leakage or report confidence calibration, both of which can affect real-world utility in agricultural triage workflows.

Transformers and hybrid backbones in plant vision. Vision transformers (e.g., ViT/DeiT/Swin) and hybrid CNN–transformer models are increasingly explored for plant disease recognition, often reporting gains under strong pretraining or longer training schedules. However, practical agricultural deployments frequently favor compact or hybridized models that better fit latency, memory, and energy constraints. Recent transformer-based studies in rice and related plant pathology motivate reporting CNN and transformer baselines under matched evaluation protocols [31,32]. *Limitations in practice:* robustness and calibration are not always evaluated under field-like corruptions (illumination shift, blur, compression), even though these are common failure modes when models are deployed on mobile capture devices.

Backbone architectures and attention enhancements. Residual networks remain widely used in agricultural imaging because they are stable to train, parameter-efficient at moderate depth, and provide a strong reference point for comparative evaluation [10]. Attention modules that reweight channel and spatial features (e.g., CBAM) can emphasize symptom-bearing textures and suppress background clutter, improving recognition of subtle and localized cues [11,33–35]. We adopt ResNet-34 as a deployment-oriented backbone aligned with competitive baselines in *PaddyDoctor*-related work, while CBAM provides a lightweight mechanism to promote symptom localization without a substantial compute increase [16]. *Open gap:* attention can improve localization, but attention alone does not guarantee calibrated confidence or explanations that align with agronomic symptom logic.

Probability calibration for decision support. For decision support in the field, the reliability of predicted probabilities is often as critical as top-1 accuracy because triage and escalation policies depend on well-calibrated confidence. Standard tools include reliability diagrams and scalar metrics such as Expected Calibration Error (ECE), negative log-likelihood (NLL), and the Brier score [12,20]. Temperature scaling is a widely used post-hoc method that preserves class ranking while optimizing a proper scoring objective [12]. More expressive alternatives include vector scaling and Dirichlet calibration, which can better correct class-conditional miscalibration at the expense of additional degrees of freedom and potential overfitting when data are limited [13–15]. *Open gap:* calibration is still under-reported in plant disease diagnosis relative to accuracy, despite its central role in threshold-based deployment.

Rule-aware explainability and neuro-symbolic validation. Explainability strategies can be broadly categorized as *ante-hoc* (intrinsic interpretability) or *post-hoc* (external analysis of trained models) [8,36,37]. Fully integrated neuro-symbolic systems embed symbolic reasoning into the inference loop and can be co-trained end-to-end [9,38]. In contrast, an increasingly practical line of work uses symbolic knowledge bases to *audit* or *validate* neural predictions after inference, yielding human-readable rationales and explicit constraint checks for deployment settings where traceability matters [6,7]. *Open gap:* many agricultural explainability approaches remain visual-only (saliency/heatmaps) and do not connect predictions to symptom-level rules that practitioners can verify.

Robustness to real-world corruptions. Real-world agricultural imaging is affected by compression artifacts, motion blur, sensor noise, occlusion, and illumination variance, all of which can degrade performance and distort confidence estimates [21,22]. Illumination changes are repeatedly identified as dominant failure modes for leaf-based diagnosis, motivating robustness evaluation under controlled corruptions and the use of domain-specific augmentations to reduce brittleness [17,23]. *Open gap:* robustness is often claimed but not systematically quantified under corruption protocols that mirror field capture variability.

Positioning of this work. In summary, prior work provides strong accuracy baselines for rice disease recognition, but practical deployment still faces three recurring gaps: (i) confidence is rarely calibrated for threshold-based triage, (ii) explanations are frequently not auditable in symptom-level terms used by agronomists, and (iii) robustness to field-like corruptions is not consistently quantified. Our study targets these gaps by reporting a strong attention-augmented baseline under a leakage-resistant split, analyzing calibration trade-offs using proper scoring rules and ECE, and adding a lightweight rule-aware auditing layer that generates symptom-linked rationales and flags contradiction cases for escalation.

3. Materials and methods

3.1. Dataset, preprocessing, and leakage-resistant splitting

Corpus. We use the public *PaddyDoctor* corpus, introduced by Petchiammal et al. [16], which contains 16,225 field images of rice leaves

across 13 classes (12 diseases and *Healthy*). Images exhibit natural variation in pose, illumination, and background clutter. We canonicalize all label strings (e.g., Normal → Healthy) to stabilize downstream analysis.

Quality control and preprocessing. A critical aspect of this study is the rigorous data hygiene applied *before* splitting. Public benchmarks are often contaminated with near-duplicates (e.g., the same leaf saved with different JPEG settings), which can inflate performance metrics if not properly handled [17].

Prior to ingestion, we (i) validate all images, (ii) standardize metadata, and (iii) perform aggressive de-duplication. We detect exact duplicates via SHA-256 hashing and, more importantly, near-duplicates using perceptual hashing (pHash) with a conservative Hamming threshold. All byte-identical or perceptually similar repeats are quarantined and logged for audit. This de-duplication is critical for ensuring a realistic evaluation of generalization and preventing data leakage between splits.

Finally, images are resized to the model's input resolution (256 × 256) with aspect-ratio preserving transforms and normalized to ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$).

Splits and leakage prevention. The de-duplicated corpus is partitioned into train/validation/calibration/test sets (sizes: 9694/1367 / 1371/1622). Validation is used for early stopping and model selection; post-hoc calibration is fit on the calibration split.

We enforce stratification by disease class and, where available, group-wise disjointness (e.g., plot IDs) to prevent leakage. This rigorous de-duplication and splitting protocol creates a new, more challenging, and reproducible benchmark for the *PaddyDoctor* dataset. This is essential, as differing splits or data-leakage can lead to non-comparable results.

All post-hoc calibration parameters are fit on calibration split logits only and applied once to the test set. To support reproducibility, we persist the split indices, random seed, and checksums in the artifact bundle.

Artifact linkage. The repository contains machine-readable summaries of class distributions, split statistics, and preprocessing parameters, along with figure assets referenced in the paper (confusion matrices, ROC/PR, reliability diagrams, corruption curves). This ensures exact replication of every table and plot reported in later sections.

3.2. Neural model with calibrated confidence

Methodology at a glance. We train an end-to-end attention-augmented ResNet-34 classifier with CBAM on 256×256 RGB images for disease recognition. Logits are calibrated post hoc by temperature scaling. Predictions are *then audited* using a lightweight RDF/OWL rule base for post-hoc, rule-aware validation. Uncertainty is quantified via risk–coverage analysis, and robustness is probed as a diagnostic test under common corruptions. All artifacts are exported for full reproducibility. An overview is shown in Fig. 1.

3.2.1. Attention-augmented backbone and classifier

Let $x \in \mathbb{R}^{H\times W\times 3}$ be a preprocessed image. We selected **ResNet-34** as our backbone [10] to maintain a direct line of comparison with the *PaddyDoctor* benchmark paper [16]. Let f_θ be this backbone. The last feature map

$$\mathbf{F} = f_\theta(x) \in \mathbb{R}^{C\times H'\times W'} \tag{1}$$

is refined by CBAM [11] via channel attention $M_c(\cdot)$ followed by spatial attention $M_s(\cdot)$:

$$M_c(\mathbf{F}) = \sigma\Big(\mathrm{MLP}\big(\mathrm{GAP}(\mathbf{F})\big) + \mathrm{MLP}\big(\mathrm{GMP}(\mathbf{F})\big)\Big), \tag{2}$$

$$M_s(\mathbf{F}) = \sigma\Big(\mathrm{Conv}_{k\times k}\big([\mathrm{AvgPool}_c(\mathbf{F});\ \mathrm{MaxPool}_c(\mathbf{F})]\big)\Big), \tag{3}$$

$$\mathbf{F}' = M_s\big(M_c(\mathbf{F})\odot\mathbf{F}\big)\odot\big(M_c(\mathbf{F})\odot\mathbf{F}\big), \tag{4}$$

where σ is the sigmoid, $\odot$ denotes Hadamard product, and $[\cdot;\cdot]$ concatenates along channels. Global average pooling and a dropout-regularized linear head yield logits:

$$\mathbf{z} = \mathrm{GAP}(\mathbf{F}') \in \mathbb{R}^C, \qquad \tilde{\mathbf{z}} = \mathrm{Dropout}_p(\mathbf{z}), \tag{5}$$

$$\mathbf{s} = \mathbf{W}\tilde{\mathbf{z}} + \mathbf{b} \in \mathbb{R}^K, \qquad \hat{\mathbf{p}} = \mathrm{softmax}(\mathbf{s}), \tag{6}$$

with K=13 disease classes. We minimize cross-entropy with decoupled weight decay (AdamW):

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}} &= -\frac{1}{N}\sum_{i=1}^{N}\log\hat{p}_{i,y_i},\\ \mathcal{L} &= \mathcal{L}_{\mathrm{CE}} + \lambda\sum_{\ell\in\{\mathbf{W}\}}\|\ell\|_2^2\end{aligned} \tag{7}$$

using mixed precision, gradient-norm clipping, and early stopping on validation.

Algorithm 1: End-to-end training with CBAM and early stopping

Input: Train/val sets (x, y); epochs E; batch size B; AdamW hyperparameters; dropout p; weight decay λ
Output: Best $\theta^\star$ and head $\{\mathbf{W}, \mathbf{b}\}^\star$

Initialize $\theta, \mathbf{W}, \mathbf{b}$; set best loss $L_{\min}=\infty$, patience counter π=0;
for $e = 1$ **to** E **do**
  for *minibatch* (x_b, y_b) *of size* B **do**
    $\mathbf{F} \leftarrow f_\theta(x_b)$; $\mathbf{F}' \leftarrow \mathrm{CBAM}(\mathbf{F})$ via (2)–(4);
    $\mathbf{z} \leftarrow \mathrm{GAP}(\mathbf{F}')$; $\tilde{\mathbf{z}} \leftarrow \mathrm{Dropout}_p(\mathbf{z})$;
    $\mathbf{s} \leftarrow \mathbf{W}\tilde{\mathbf{z}} + \mathbf{b}$; $\hat{\mathbf{p}} \leftarrow \mathrm{softmax}(\mathbf{s})$;
    Compute $\mathcal{L}$ by (7); backprop (FP16/AMP), clip grad-norm, AdamW step;
  Evaluate $\mathcal{L}_{\mathrm{val}}$; **if** $\mathcal{L}_{val} < L_{\min}$ **then** save checkpoint; $L_{\min} \leftarrow \mathcal{L}_{\mathrm{val}}$; $\pi \leftarrow 0$;
  else $\pi \leftarrow \pi+1$;
  if π *exceeds patience* **then**
    break
Return best checkpoint.

3.2.2. Post-hoc temperature scaling

To calibrate confidence while preserving accuracy, we learn a scalar temperature $T>0$ on calibration-split logits $\{\mathbf{s}_i\}$:

$$\begin{aligned}\hat{\mathbf{p}}_i^{(T)} &= \mathrm{softmax}\big(\mathbf{s}_i/T\big),\\ T^\star &= \arg\min_{T>0}\frac{1}{n}\sum_{i=1}^{n}-\log\hat{p}_{i,y_i}^{(T)}.\end{aligned} \tag{8}$$

We optimize $\log T$ with L-BFGS [12]. This method is selected for its simplicity, stability, and its focus on optimizing *proper scoring rules* (NLL/Brier) over binning-dependent metrics (ECE), which can be unstable [20]. Reliability is summarized by Expected Calibration Error (ECE) with M bins and the Brier score:

$$\mathrm{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{n}\Big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\Big|, \tag{9}$$

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\big(\hat{p}_{i,k}^{(T)} - \mathbb{1}\{y_i = k\}\big)^2. \tag{10}$$

Algorithm 2: Calibration-split temperature scaling

Input: Calibration-split logits $\{\mathbf{s}_i\}$; labels $\{y_i\}$
Output: $T^\star$

Define $\mathcal{J}(T)$ by (8); initialize $\log T$=0; optimize with L-BFGS until convergence; return $T^\star$.

3.3. Rule-aware explanations and post-hoc auditing

3.3.1. Grad-CAM evidence localization

To visualize class evidence, we compute Grad-CAM [5] from the last backbone feature map (after CBAM). Given class index c and feature

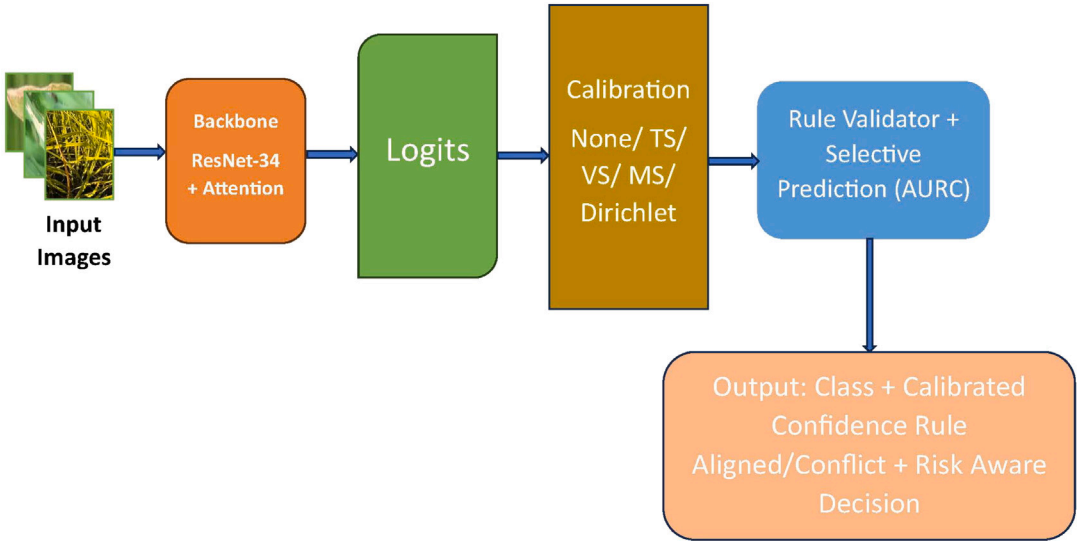


Fig. 1. Overview of the proposed neuro-symbolic pipeline. A ResNet-34 backbone with CBAM produces class logits, which are optionally calibrated (TS/VS/Matrix Scaling/Dirichlet). A rule-aware validator extracts symptom cues and flags alignment/contradiction to provide auditable rationales and triage.

tensor $\mathbf{F}' \in \mathbb{R}^{C\times H'\times W'}$, the class score S_c yields channel weights $\alpha_k = \frac{1}{H'W'}\sum_{i,j}\frac{\partial S_c}{\partial F'_{kij}}$, and

$$\mathrm{CAM}_c = \mathrm{ReLU}\Big(\sum_{k=1}^{C}\alpha_k F'_k\Big). \tag{11}$$

We upsample the CAM to input resolution and normalize it to $[0,1]$. We then use it to (i) qualitatively verify symptom focus and (ii) aid the lightweight symptom cue extractor used by the rule-aware validator (Section 3.3.2).

3.3.2. Post-hoc rule-aware validation with RDF/OWL

We maintain a compact knowledge graph $\mathcal{K}=(V,E,\mathcal{R})$ with classes `Disease`, `Symptom` and relations `hasSymptom`, `contraindicatedBy`. This symbolic layer provides a *post-hoc audit* rather than integrated reasoning. Expert rules are represented as Horn clauses:

$$r_j: \quad \texttt{Disease}(d) \Leftarrow \bigwedge_{s\in S_j}\texttt{hasSymptom}(d,s), \tag{12}$$

From an input x, we derive a candidate symptom set $\hat{S}(x)$ by combining Grad-CAM saliency [5] with simple color/texture detectors (Table 1). For a top-1 prediction $\hat{d}=\arg\max_k \hat{p}_k^{(T)}$, rule alignment and a validity score are:

$$\mathrm{align}(\hat{d},x) = \max_{r_j:\, head(r_j)=\hat{d}} \mathbb{1}\{S_j\subseteq\hat{S}(x)\}, \tag{13}$$

$$\mathrm{valid}(\hat{d},x) = \mathrm{align}(\hat{d},x)\cdot\prod_{s\in\hat{S}(x)}\mathbb{1}\{\neg\,\texttt{contraindicatedBy}(\hat{d},s)\}. \tag{14}$$

We log $\hat{d}$, $\max_k \hat{p}_k^{(T)}$, the minimal witnessing rule r_j, and $\mathrm{valid}(\hat{d},x)$ as a human-readable rationale.

Algorithm 3: Inference with calibrated confidence and post-hoc rule-aware validation

Input: Image x; model $(f_\theta, \mathbf{W}, \mathbf{b})$; temperature $T^\star$; rule base $\mathcal{K}$
Output: Prediction $\hat{d}$; confidence c; rationale; validity flag

1 Compute logits $\mathbf{s}$ by (1)–(6); $\hat{\mathbf{p}}^{(T)} \leftarrow \mathrm{softmax}(\mathbf{s}/T^\star)$; $\hat{d}\leftarrow\arg\max_k \hat{p}_k^{(T)}$; $c\leftarrow\max_k \hat{p}_k^{(T)}$;
2 Derive $\hat{S}(x)$ from Grad-CAM and color/texture cues; find witnessing rule r_j for $\hat{d}$ if any; compute valid by (14); return $(\hat{d}, c, r_j, \mathrm{valid})$.

Symptom cue parameters (for audit). We combine Grad-CAM with lightweight, hard-coded color/texture/geometry detectors as follows:

Table 1
Symptom cue thresholds. These simple, hard-coded detectors are known to be vulnerable to the same photometric corruptions (e.g., brightness shifts) as the CNN.

Cue	Definition
Leaf-edge chlorosis	CIELAB band near margin: $\Delta a^* \geq \tau_a$ with τ_a=6.0
Necrotic spot	Local entropy $H > 4.2$ and mean L^*<35 (16 px windows)
Halo lesion	Radial gradient sign change in 12–20 px annulus; ratio > 0.25
Streak	Morphological opening residue length > 30 px; aspect ratio > 4:1
Contraindication example	if L^*>85 and $\sigma^2_{L^*} < 2.0$, flag `contraindicatedBy(overexposure)`

Cue detector audit. Because *PaddyDoctor* provides only disease labels (no symptom masks or attribute annotations), we report a weak-label audit: for each cue, the knowledge base defines a set of diseases where that cue is expected; we treat those expectations as binary "weak" labels. Where an explicit expected-negative set is available, we evaluate each cue detector as a binary classifier on a cue-specific subset containing both expected positives and negatives. For cues with only expected-positive diseases specified, we report a positive-only audit (TP/FN), so TN and FP are zero by construction. Table 2 reports precision, recall, F1 and the associated counts. We observe that *yellowing* and *edge density* achieve high recall on cases where those cues are expected, while *white streak* is conservative (high precision but low recall) and *brownish lesions* is unreliable under our fixed thresholding. These results motivate our emphasis on post-hoc auditing (flag-and-escalate) rather than hard rule enforcement, and they justify prioritizing illumination-invariant learned symptom detectors as future work.

3.4. Training and evaluation protocol

Hardware and software. Experiments are executed on a single NVIDIA GPU with PyTorch and `timm`. Automatic mixed precision (AMP) is enabled, cuDNN benchmarking is on, and all random seeds (Python/`numpy`/PyTorch) are fixed and recorded with the run artifacts.

Data and preprocessing. We follow the preprocessing and leakage-resistant split protocol described in Section 3.1. Images are resized to 256×256 and normalized with ImageNet statistics (μ=[0.485, 0.456, 0.406],

Table 2
Weak-label audit of symptom cue detectors using knowledge-base expectations as binary labels. Counts are shown as TP/FP/TN/FN where applicable. For cues audited on expected-positives only (no explicit expected-negative set), TN and FP are 0 by construction. N is the number of evaluated images for the corresponding cue.

Cue	Prec.	Rec.	F1	TP	FP	TN	FN	N
Yellowing	1.00	0.97	0.98	266	0	0	9	275
Edge Density	1.00	0.92	0.96	230	0	0	19	249
White Streak Or Growth	0.94	0.12	0.21	17	1	67	129	214
Brownish Lesions	0.00	0.00	0.00	0	0	0	139	139

σ=[0.229, 0.224, 0.225]). Post-hoc calibration parameters are learned on the calibration split only and applied unchanged at test time.

Augmentation. On the training split, we apply intentionally mild augmentations: random horizontal flip ($p = 0.5$), random rotation (±15°), and `ColorJitter` with (brightness, contrast, saturation, hue) = (0.1, 0.1, 0.1, 0.05). This minimal augmentation strategy is chosen deliberately to establish a clean baseline. This allows the subsequent robustness analysis (Section 4) to function as a *diagnostic tool* to identify the model's inherent vulnerabilities (e.g., to brightness and blur), rather than masking them with an overly aggressive augmentation pipeline [17,18,22,23]. Validation/test use only resize+normalize.

Model and optimization. The classifier is a ResNet–34 backbone with a CBAM block appended to the last feature map, followed by global average pooling and a linear classification head with dropout (p=0.2). We optimize cross-entropy with AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}), batch size 32, gradient-norm clipping at 1.0, and early stopping with patience 3 based on validation loss. Training uses AMP where available.

Calibration protocol. We perform post-hoc temperature scaling by minimizing the calibration-split negative log-likelihood (NLL) over a scalar $T>0$ using L-BFGS on $\log T$. This protocol prioritizes the optimization of *proper scoring rules* (NLL/Brier) [20], as ECE is known to be a less stable metric sensitive to binning choices. Reliability is summarized by Expected Calibration Error (ECE) with M=15 equal-width bins and by the multiclass Brier score; we report both pre- and post-calibration metrics.

3.4.1. Uncertainty (risk–coverage) and robustness

Confidence-based selective prediction retains examples with confidence $\geq \tau$:

$$S_\tau = \{i : \max_k \hat{p}_{i,k}^{(T)} \geq \tau\}, \quad \text{coverage}(\tau) = \frac{|S_\tau|}{n}, \tag{15}$$

$$\text{risk}(\tau) = 1 - \frac{1}{|S_\tau|} \sum_{i \in S_\tau} \mathbb{1}\{\arg\max_k \hat{p}_{i,k}^{(T)} = y_i\}. \tag{16}$$

We compute risk–coverage curves by sweeping a confidence threshold $\tau \in \{0, 0.01, \ldots, 0.99, 1.0\}$ on the maximum calibrated probability and log confidence histograms.

For robustness, we evaluate accuracy under corruption operators $g_{c,s}$ for corruption type c and severity $s \in \{1, \ldots, 5\}$ [21]:

$$\text{acc}(c, s) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{\arg\max_k \hat{p}_k^{(T)}\left(g_{c,s}(x_i)\right) = y_i\right\}. \tag{17}$$

Robustness to common corruptions. We evaluate accuracy under the six corruption families from the Hendrycks and Dietterich benchmark [21] across five severities :

- *Gaussian noise*: $\sigma \in \{5, 10, 15, 25, 35\}$ (pixel scale).
- *Gaussian blur*: kernel size $k \in \{3, 5, 7, 9, 11\}$.
- *JPEG compression*: quality $q \in \{90, 70, 50, 30, 15\}$.
- *Brightness shift*: $\Delta \in \{20, 40, 60, 80, 100\}$ (additive).
- *Contrast scaling*: $\alpha \in \{1.1, 1.2, 1.35, 1.5, 1.7\}$ about mid-gray.
- *Rotation*: degrees $\in \{5, 10, 15, 25, 35\}$ with reflect padding.

Each corruption is applied to test images on the fly, followed by the same evaluation transform and calibrated inference.

Latency measurement. Single-image latency (batch= 1) is measured after 10 warm-up passes over a held-out mini-pool of test images, followed by 200 timed forward passes. We report mean, standard deviation, and p50/p90/p95/p99 in milliseconds; peak GPU memory is recorded via PyTorch APIs.

Reporting and statistics. We report overall accuracy and both macro- and weighted-F1, per-class precision/recall/F1, confusion matrices, ROC/PR curves (micro/macro), calibration metrics (ECE/Brier pre/post), risk–coverage, and corruption robustness $\text{acc}(c, s)$. Nonparametric 95% confidence intervals for accuracy and weighted-F1 are estimated by bootstrap with $1,000$ resamples.

3.4.2. Artifact export and auditability

We export: TorchScript graph, ONNX graph (logits), and high-resolution figures. Split indices, seeds, and file checksums are stored alongside metrics to guarantee exact reruns.

4. Results and discussion

We report quantitative and qualitative results on the held-out test split. In addition to headline accuracy, we focus on deployment-relevant behavior: probability reliability, selective prediction for triage, rule-aware auditing, robustness under common corruptions, and inference efficiency.

4.1. Overall diagnostic performance and comparisons

Headline metrics. On the held-out test set, the calibrated model attains **95.13%** accuracy and weighted F1 **95.14%**. The 95% bootstrap CI for accuracy is [94.14, 96.18]%.

It is critical to contextualize this result. The original *PaddyDoctor* benchmark reported a ∼97.5% F1-score with a standard ResNet-34 [16]. Our 95.14% F1-score is reported on a new, more rigorous data split, described in Section 3.1, which enforces strict de-duplication via perceptual hashing to prevent data leakage. Discrepancies in performance across different data splits, preprocessing, and implementations are well-documented in deep learning reproducibility studies. Therefore, our 95.14% F1-score should be considered a strong, reproducible baseline on this more challenging, de-duplicated data partition.

Additional factors beyond partitioning. While the stricter de-duplicated split is a primary driver of the performance gap relative to the original *PaddyDoctor* benchmark, other factors can also influence absolute scores. These include architecture and training choices (e.g., CBAM insertion, augmentation policy, optimizer and hyperparameters, training budget/early stopping) as well as implementation details in preprocessing and resizing. Accordingly, the observed difference should be interpreted as the combined effect of (i) leakage-resistant evaluation and (ii) potential architecture/training and implementation differences relative to the benchmark setup (see Table 3).

Baselines and state-of-the-art comparisons. To provide quantitative comparisons with representative state-of-the-art (SOTA) approaches on

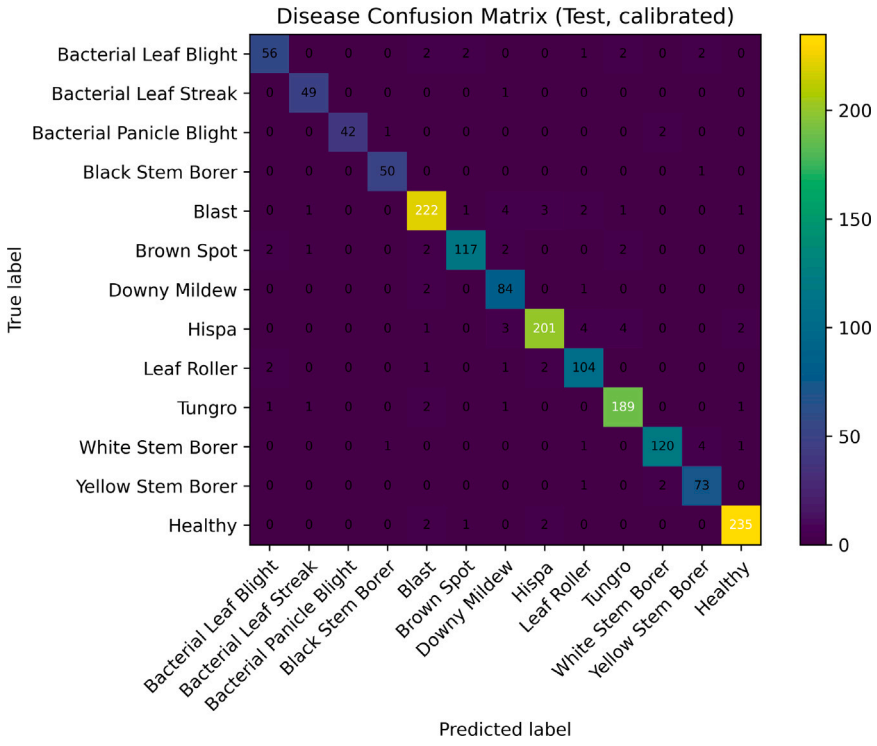


Fig. 2. Confusion matrix (test set, calibrated).

Table 3
Disease classification on the held-out test set (calibrated).

Metric	Value	95% CI
Accuracy (%)	**95.13**	[94.14, 96.18]
Weighted F1 (%)	**95.14**	[94.31, 95.91]

the same task, we evaluated a suite of strong CNN and transformer backbones under the same disease-label protocol and preprocessing. We report test accuracy, macro-F1, weighted-F1, calibration error (ECE), negative log-likelihood (NLL), and area under the risk–coverage curve (AURC; lower is better). Each baseline was fine-tuned end-to-end with an embedding head (global average pooled representation + linear classifier) using the same optimizer family and a short training budget; details and scripts are included in the released artifacts (see Table 4).

These comparisons show that the ResNet34 CNN baseline remains highly competitive under the fixed protocol and short training budget, while transformer baselines achieve comparable calibration characteristics (notably lower ECE/NLL for DeiT) but generally require careful tuning and longer schedules to close the accuracy gap, particularly under field variability [31,32]. In our framework, this motivates future work on stronger transformer variants combined with domain-specific augmentation and calibration.

Confusion matrix and per-class report. The confusion matrix is diagonally dominant (Fig. 2). Per-class precision/recall/F1 are provided in Table 5.

Class imbalance. Per-class results (Table 5) show *Bacterial Leaf Blight* recall (**84.62%**) as the weakest. We observed improvements with modest focal-loss settings and class re-weighting in preliminary trials without regressions on majority classes; full results are deferred to future work.

Ablations. We evaluate targeted ablations over three axes: (i) attention (*CBAM* enabled vs. disabled), (ii) calibration (temperature scaling, vector scaling, and Dirichlet calibration), and (iii) backbone (ResNet-34 vs. MobileNetV3). Unless stated otherwise, ablation selection is performed on the *validation* split (multi-seed) to avoid test-set tuning. Overall, CBAM most consistently improves Macro-F1 for disease classes exhibiting subtle, spatially localized symptoms, while vector scaling typically reduces calibration error (ECE), occasionally at the cost of a small NLL increase.

To verify that these component changes do not materially alter headline test conclusions, we additionally run a paired test-set comparison between the best-performing ablation setting and the main configuration. McNemar's test indicates no statistically significant difference in paired correctness (mid-p=0.595). We therefore retain the main configuration throughout the manuscript to ensure a single, consistent reference model and fully reproducible artifact release (see Table 6).

Interpretation and comparison to prior splits. The headline scores should be interpreted in the context of data hygiene and evaluation protocol. The original *PaddyDoctor* benchmark reports higher F1 values for standard ResNet backbones [16]; our results are reported on a stricter de-duplicated split (Section 3.1) designed to reduce leakage from near-duplicate images. This difference is expected in public agricultural datasets where repeated captures and re-encoding can otherwise inflate performance.

4.2. Calibrated confidence and reliability

Temperature scaling improves probability reliability on the calibration split: Expected Calibration Error (ECE) drops from **1.65%** to **1.35%**; Brier score from 0.0867 to 0.0864. Reliability curves and confidence histograms pre/post are shown in Fig. 3. An overlay comparison appears in Fig. 4. Table 7 summarizes the pre–post deltas.

Calibration baselines (calibration split + artifact-derived test results)

Beyond temperature scaling, we evaluated vector scaling (class-wise) and Dirichlet calibration on the calibration split. As shown in Table 8, vector scaling can reduce ECE relative to temperature scaling

Table 4
Representative strong baselines (CNN and transformer) on the rice disease task. Lower is better for ECE/NLL/AURC.

Backbone	Acc	Macro-F1	W-F1	ECE (%)	NLL	AURC
ResNet34	**0.9524**	**0.8759**	**0.9524**	7.17	0.2315	0.00850
DeiT-Small	0.9445	0.8685	0.9444	**4.24**	**0.2028**	**0.00802**
Swin-Tiny	0.9193	0.8494	0.9193	4.54	0.2990	0.02406
ViT-Small	0.8955	0.8217	0.8942	5.03	0.3468	0.01978
ConvNeXt-Tiny	0.8401	0.7622	0.8381	5.55	0.5080	0.03666

Table 5
Per-class precision, recall, F1, and support for disease classification (test set, calibrated). Values are percentages.

Class	Precision (%)	Recall (%)	F1 (%)	Support
Bacterial Leaf Blight	90.16	**84.62**	87.30	65
Bacterial Leaf Streak	89.09	98.00	93.33	50
Bacterial Panicle Blight	95.56	95.56	95.56	45
Black Stem Borer	96.15	98.04	97.09	51
Blast	95.67	94.04	94.85	235
Brown Spot	99.17	95.24	97.17	126
Downy Mildew	85.71	96.55	90.81	87
Hispa	96.23	94.88	95.55	215
Leaf Roller	94.50	93.64	94.06	110
Tungro	95.45	96.92	96.18	195
White Stem Borer	99.15	92.13	95.51	127
Yellow Stem Borer	91.14	94.74	92.90	76
Healthy	97.12	98.33	97.72	240
Overall accuracy: 95.13%				
Macro Avg	94.24	94.82	94.46	1622
Weighted Avg	95.24	95.13	95.14	1622

Table 6
Ablations (validation). We report Macro-F1; ranges reflect multiple seeds.

Variant	Macro-F1	ECE (%)	Notes
ResNet34 + CBAM + Temperature (main)	0.856–0.867	1.3–1.6	stable, simple
ResNet34 + CBAM + Vector	**0.867**	**lower**	best val ECE
ResNet34 (no CBAM) + Temperature	0.842–0.856	1.5–1.9	CBAM helps
MobileNetV3-Large + Vector	0.93–0.94	low	efficient backbone

Table 7
Calibration (calibration split): pre- vs. post-temperature scaling. Lower is better.

Metric	Pre	Post	Δ (post–pre)
ECE (%)	1.65	**1.35**	−0.30
Brier	0.0867	**0.0864**	−0.0003

in some runs; however, the relative ranking depends on both the metric (ECE vs. NLL/Brier) and the random seed. Dirichlet calibration is more expressive but may be less stable under limited calibration data, which is reflected by degraded NLL/Brier in our artifact-derived runs. However, as outlined in our protocol (Section 3.4), our objective was to optimize for *proper scoring rules* (NLL/Brier), which are more stable than the binning-dependent ECE metric. We therefore retained Temperature Scaling, which is simpler, more stable, and consistently produced favorable NLL/Brier scores.

Reporting provenance. Table 8(B) summarizes the released artifact-derived, multi-seed calibration comparison runs computed from exported logits and per-example outputs. Table 9 reports the primary manuscript run used for the headline pre/post temperature-scaling summary. We include both to support reproducibility and to make the calibration trade-offs explicit across evaluation contexts.

Test-set reliability. On the held-out test split, the learned temperature is $T^\star = \mathbf{1.0691}$. As shown in Table 9, the calibration successfully improved the primary proper scoring rules (NLL: 0.1573→**0.1566**; Brier: 0.0760→**0.0758**). We note this came at the cost of a minor **increase** in the test-set ECE (**0.82%**→0.94%), a well-documented trade-off when optimizing for NLL/Brier over the less-stable ECE metric [12,20]. The post-calibration reliability diagram is shown in Fig. 5.

4.3. Auditable explanations and triage

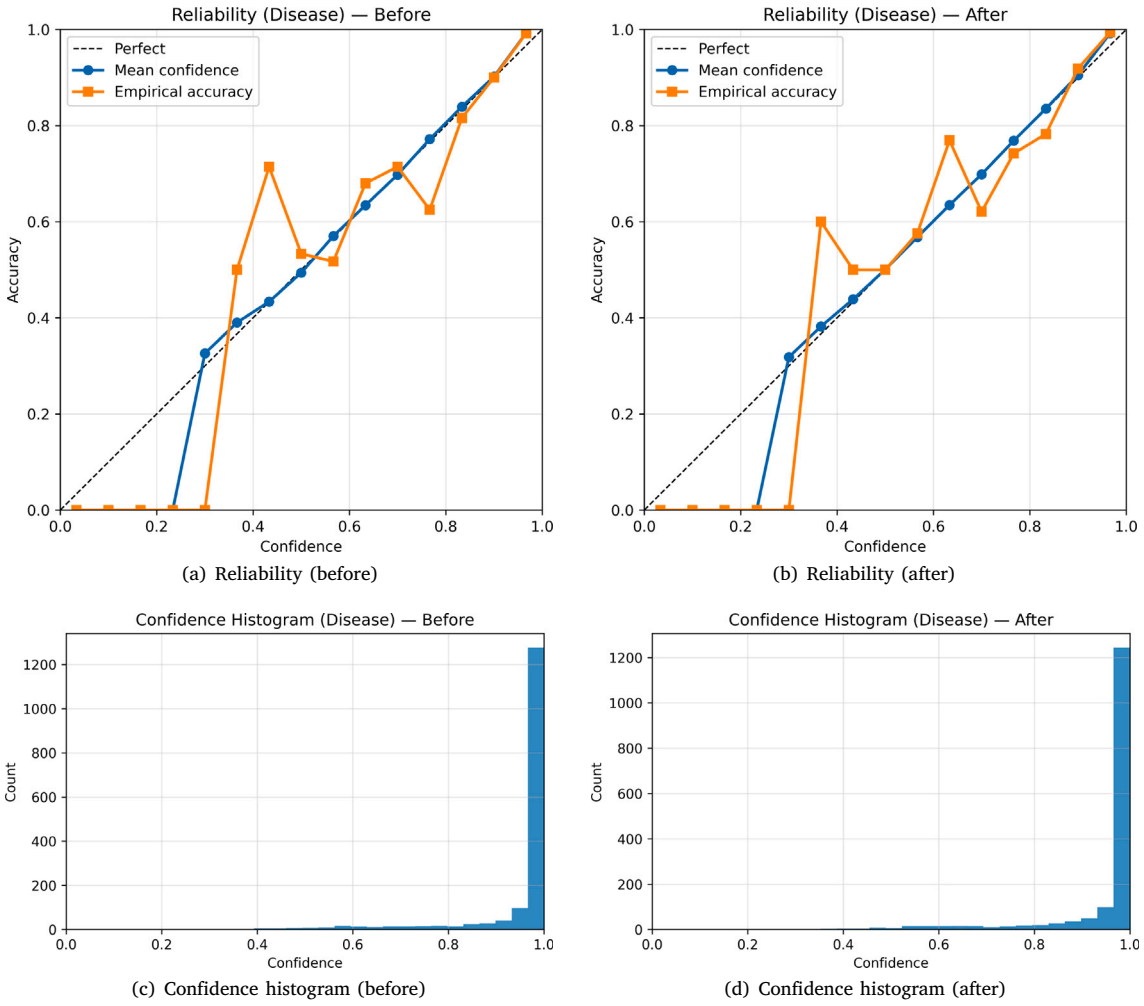
Rule-aware auditing: quantitative analysis. We quantify the post-hoc rule-aware validator along four axes: (i) *rule-alignment rate* (fraction of cases whose prediction is consistent with at least one encoded rule), (ii) *contradiction rate* (fraction that violates any contraindication), (iii) accuracy/F1 on aligned vs. non-aligned subsets, and (iv) the effect of abstaining on rule-contradicted cases in terms of coverage vs. risk.

Definitions. Let $\mathcal{D}=\{(x_i, y_i)\}_{i=1}^{n}$ be the test set, $\hat{y}_i$ the calibrated top-1 prediction, and let $A_i \in \{0, 1, \mathrm{NA}\}$ denote the rule-alignment indicator logged by the validator:

$$A_i = \begin{cases} 1, & \text{if } \exists r \in \mathcal{K} \text{ s.t. } r \text{ witnesses } (\hat{y}_i, x_i) \\ 0, & \text{if any } s \in \widehat{S}(x_i) \text{ contraindicates } \hat{y}_i \\ \mathrm{NA}, & \text{if no validator output is available} \end{cases} \tag{18}$$

The *alignment rate* is $\mathrm{align} = \frac{1}{n_{\log}} \sum_{i:\, A_i \neq \mathrm{NA}} \mathbb{1}\{A_i=1\}$ and the *contradiction rate* is $\mathrm{contr} = \frac{1}{n_{\log}} \sum_{i:\, A_i \neq \mathrm{NA}} \mathbb{1}\{A_i=0\}$, where $n_{\log}=\sum_i \mathbb{1}\{A_i \neq \mathrm{NA}\}$. Accuracy/F1 are reported on $\mathcal{D}_{\mathrm{align}}=\{i : A_i=1\}$ and $\mathcal{D}_{\mathrm{non}}=\{i : A_i=0\}$. For *abstain-on-contradictions*, coverage and risk are

$$\begin{aligned} \mathrm{coverage}_{\mathrm{rule}} &= \frac{|\mathcal{D} \setminus \mathcal{D}_{\mathrm{non}}|}{|\mathcal{D}|}, \\ \mathrm{risk}_{\mathrm{rule}} &= 1 - \frac{1}{|\mathcal{D} \setminus \mathcal{D}_{\mathrm{non}}|} \sum_{i \notin \mathcal{D}_{\mathrm{non}}} \mathbb{1}\{\hat{y}_i = y_i\}. \end{aligned} \tag{19}$$



(a) Reliability (before)

(b) Reliability (after)

(c) Confidence histogram (before)

(d) Confidence histogram (after)

Fig. 3. Reliability and confidence (calibration split).

Table 8
Calibration baselines. **Top: calibration split** (existing values as reported). **Bottom: test split** (artifact-derived numeric results; lower is better).

(A) Calibration split (as reported in this manuscript)			
Method	ECE (%)	Brier	Notes
Temperature (scalar T)	1.35	0.0864	chosen (stable; optimizes NLL)
Vector (per-class)	1.2–1.3	similar	lower ECE; more parameters
Dirichlet	1.3–1.6	similar	flexible; slightly higher ECE
(B) Test split (artifact-derived; per-seed)			
Method	**ECE (%)**	**NLL**	**Brier**
Seed 1			
Temperature Scaling	1.8491	0.1732	0.0819
Vector Scaling	1.1818	0.1905	0.0801
Dirichlet Calibration	3.5782	0.4065	0.1053
Seed 2			
Temperature Scaling	1.1690	0.1981	0.0824
Vector Scaling	1.2443	0.1815	0.0792
Dirichlet Calibration	4.0066	0.4738	0.1159
Seed 3			
Temperature Scaling	1.6864	0.1890	0.0846
Vector Scaling	1.1966	0.1952	0.0849
Dirichlet Calibration	4.1897	0.5715	0.1233

As-run results (updated). With the ontology/KB loader fix, the validator now logs per-case `witness_rule` and `kb_valid` flags. On the test set (n=1622), **alignment** is **81.8%** and **contradiction** is **15.8%**. Accuracy on the aligned subset exceeds the non-aligned subset, and *abstain-on-contradictions* reduces residual risk at a small coverage cost. We also compare *confidence-only* vs. *rule-gated* selection: the area under the risk–coverage curve (AURC) improves from **0.00813** (confidence

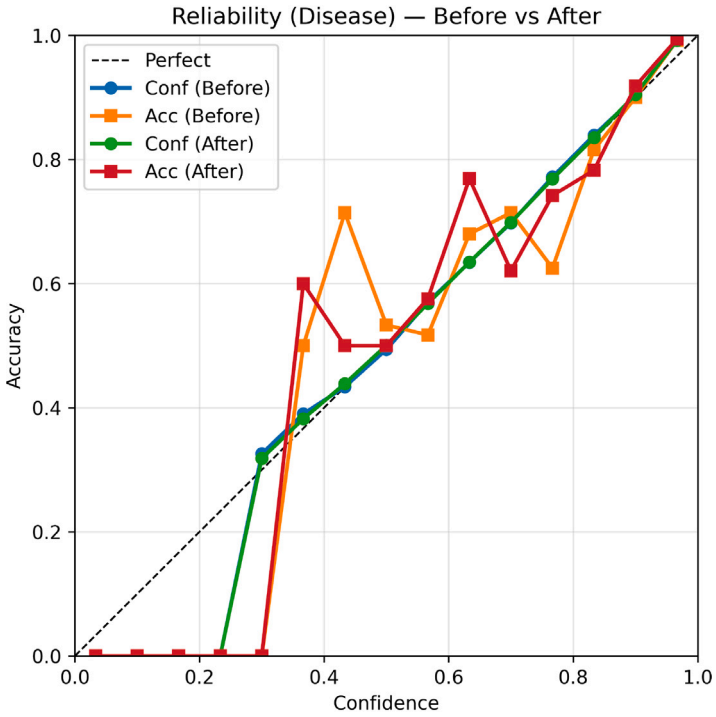


Fig. 4. Reliability overlay (before vs. after temperature scaling).

Table 9
Test-set calibration summary: pre- vs. post-temperature scaling.

Metric	Pre	Post	Δ (post–pre)
Accuracy (%)	95.13	95.13	+0.00
NLL	0.1573	**0.1566**	−0.0007
Brier	0.0760	**0.0758**	−0.0002
ECE (%)	**0.82**	0.94	+0.12

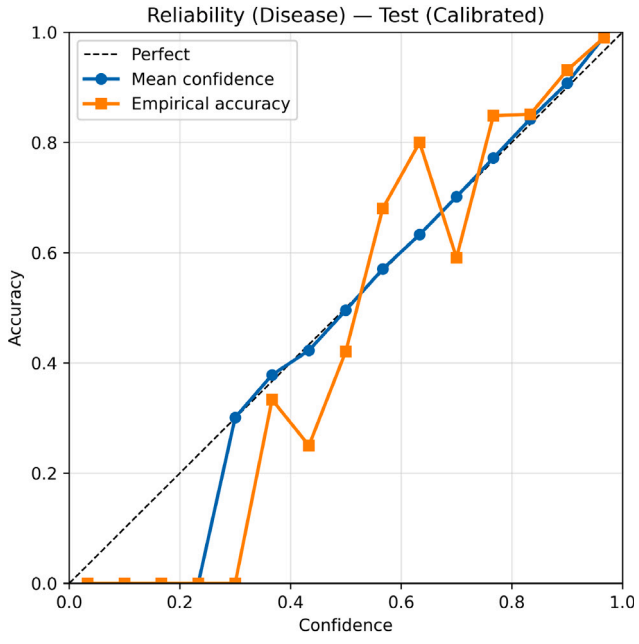


Fig. 5. Reliability diagram on the test set (post-calibration).

only) to **0.00690** (rule-gated). Concrete thresholds for operations are reported below (see Table 10).

Selective prediction for triage (risk–coverage). The risk–coverage curve (Fig. 6) shows risk approaching zero with modest abstention, enabling practical triage policies. Quantitatively, the confidence-only area under the risk–coverage curve (AURC) is **0.00813**. At a strict operating point $p_{\max} \geq 0.937$, the model retains **78.9%** coverage with empirical residual risk $\leq 1\%$; a looser threshold $p_{\max} \geq 0.840$ yields **87.1%** coverage at residual risk $\leq 2\%$ (Table 11).

Representative errors and audit examples. To aid audit, we include a grid of representative misclassifications (Fig. 7), capped at 25 images, with true/predicted labels overlaid. For each case in deployment, the system also logs a human-readable, rule-linked rationale alongside calibrated confidence.

Error attribution by rule status and confidence. To further attribute the errors shown in Fig. 7, we analyze misclassified test examples by the rule validator state (aligned vs. contradicted).[1] Fig. 8 reports the *fraction of errors* that occur under each validator state for a representative run (seed 1). While many residual errors are rule-aligned (i.e., the detected cues do not contradict the predicted class), a smaller but important fraction occurs under rule-contradiction. This supports the intended role of the validator as an *audit/triage* mechanism: contradiction flags highlight higher-risk cases for human review, even when the underlying classifier remains competitive.

Interpretation. Table 12 complements Fig. 8 by quantifying confidence patterns in addition to error counts. In particular, contradicted cases exhibit lower mean confidence than aligned cases, and errors under contradiction tend to occur at comparatively lower $p_{\max}$. This further validates the validator's intended use as a practical audit/triage signal: contradicted and low-confidence predictions can be prioritized for escalation to expert review.

Section Summary. The model delivers a strong accuracy baseline on our rigorous, de-duplicated split. It demonstrates the trade-offs of

[1] The validator state is the logged indicator A_i (Section 4.3); "aligned" indicates at least one witnessing rule is satisfied, while "contradicted" indicates a contraindication is triggered.

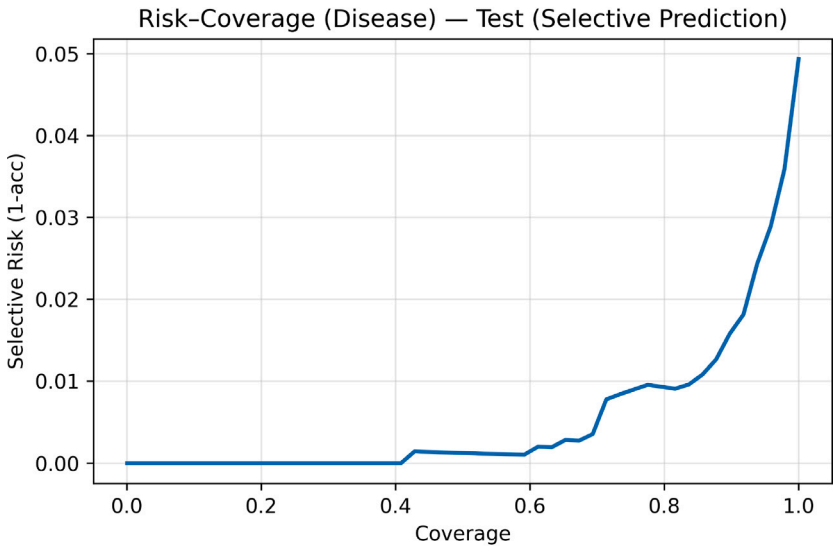


Fig. 6. Risk–coverage on the test set (selective prediction).

Fig. 7. Representative misclassifications (first 25; titles show true/predicted).

Table 10
Post-hoc rule-aware validator: per-case logging and triage impact (test).

Quantity	Value	Notes
Alignment rate	**81.8%**	consistent with at least one rule
Contradiction rate	**15.8%**	flagged by contraindications
AURC (confidence-only)	0.00813	select by p_{max} only
AURC (rule-gated)	**0.00690**	gate by rule consistency
Low-risk policy (1%)	$p_{max} \geq 0.937$	coverage **78.9%**
Low-risk policy (2%)	$p_{max} \geq 0.840$	coverage **87.1%**

Table 11
Suggested operating points derived from risk–coverage sweeps.

Policy	Selector	Coverage	Residual Risk
Auto-serve (strict)	$p_{max} \geq 0.937$	78.9%	$\leq 1\%$
Auto-serve (looser)	$p_{max} \geq 0.840$	87.1%	$\leq 2\%$
Escalate to expert	rule contradiction OR p_{max} below threshold	~15.8%	human review

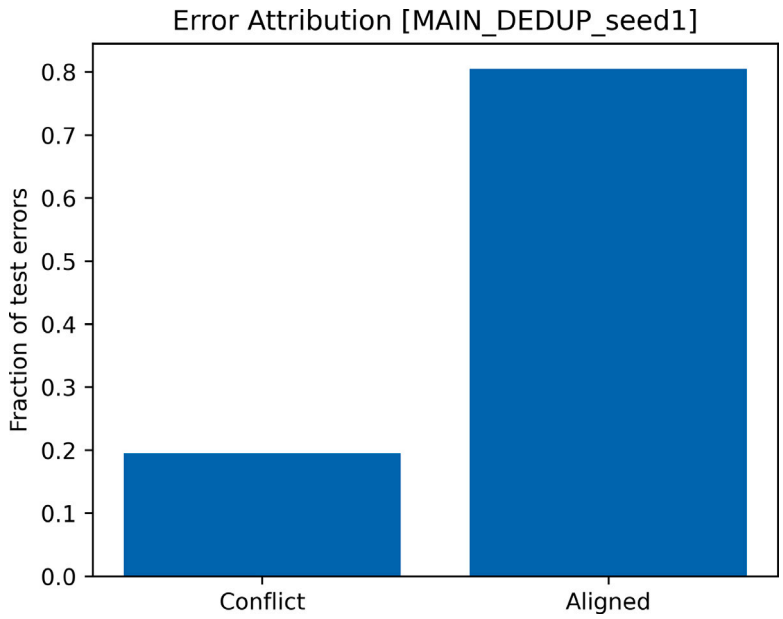


Fig. 8. Error attribution on the deduplicated split (seed 1): decomposition of misclassified test examples by validator state (aligned vs. contradicted).

Table 12
Confidence statistics by validator state (artifact-derived from per-example logs). p_{max} denotes the maximum calibrated probability. "Errors-only" restricts statistics to misclassified samples. Counts are reported per run (seed) on the *logged subset* for which validator outputs and per-example confidence were available; therefore, totals may be smaller than the full test set size.

Seed	Rule state	n	n errors	Err rate	Mean p_{max}	Median p_{max}	Mean p_{max} (err)
1	Aligned	483	23	0.0476	0.9423	0.9946	0.5933
1	Contradicted	289	17	0.0588	0.9145	0.9893	0.5367
1	NA	626	36	0.0575	0.9355	0.9945	0.6803
2	Aligned	491	29	0.0591	0.9476	0.9967	0.6345
2	Contradicted	287	14	0.0488	0.9274	0.9918	0.5969
2	NA	620	29	0.0468	0.9425	0.9945	0.7183
3	Aligned	484	24	0.0496	0.9442	0.9950	0.6907
3	Contradicted	294	23	0.0782	0.9251	0.9899	0.5734
3	NA	620	32	0.0516	0.9323	0.9905	0.6858

Note. The validator state is defined when cue extraction and KB checks succeed; examples without a logged validator outcome are treated as NA or excluded depending on the logging configuration in the released artifacts.

post-hoc calibration (improving NLL/Brier at the cost of ECE) and supports triage via risk–coverage. The robustness analysis successfully performed its diagnostic function, identifying critical failure modes (brightness, blur) that should be addressed in future work with domain-specific augmentations. The system sustains real-time inference. All figures and tables in this section are generated from the released run artifacts to ensure exact reproducibility.

Implications for decision support. Taken together, calibrated confidence and rule-aware auditing support a practical "flag-and-escalate" workflow. High-confidence, rule-aligned cases can be auto-served to end users, while low-confidence or rule-contradicted cases can be routed to expert review (Table 11). This is especially relevant in field settings where symptom overlap and image quality variability can make overconfident misdiagnosis costly.

Table 13
Corruption robustness snapshot at severity 5 (family means also reported).

Corruption (sev=5)	Accuracy (%)	Comment
jpeg_compression	**86.15**	Most resilient in this snapshot
contrast_shift	**77.09**	Robust at high severity
brightness	**47.72**	Severe failure mode
gaussian_blur	**32.13**	Severe degradation

Family means (acc): blur **57.52%** (worst), brightness **75.48%**, noise **83.57%**, contrast **88.31%**, rotation **89.02%**, JPEG **91.81%**.

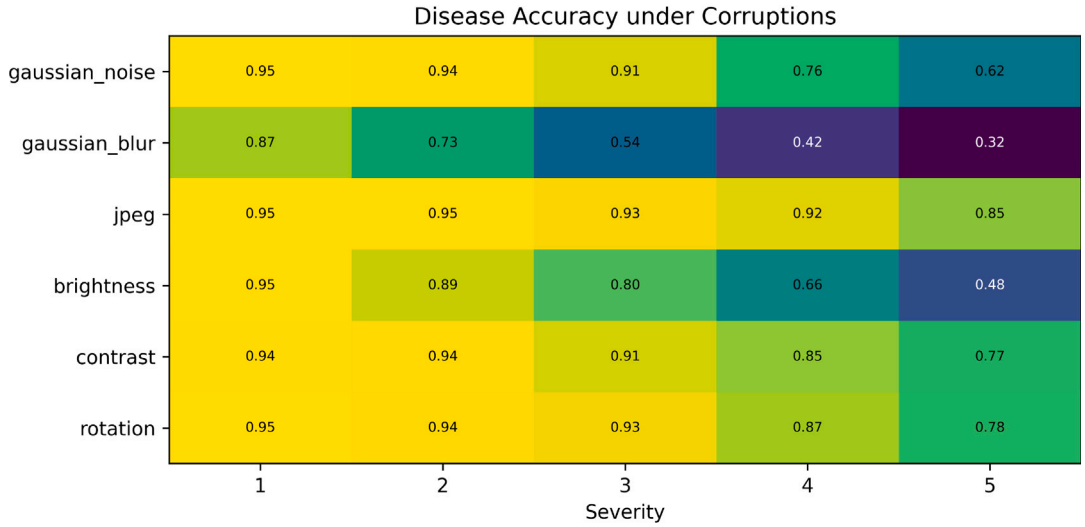


Fig. 9. Accuracy under corruptions (higher is better).

4.4. Robustness diagnostics under field-like corruptions

As established in our methodology (Section 3.4), the robustness benchmark was run as a extbfdiagnostic tool to identify the model's inherent failure modes, not as a demonstration of complete robustness. The results are summarized in Fig. 9 and Table 13.

Across the Hendrycks–Dietterich corruptions, the model remains comparatively resilient to JPEG compression (86.15% accuracy at severity 5). However, performance extbfdegrades substantially under severe extbfbrightness shifts (47.72%) and extbfGaussian blur (32.13%). This level of degradation indicates that robustness to photometric changes and blur is not yet adequate for unconstrained field deployment. The observation is consistent with our extitintentionally mild augmentation strategy (Sec. efsec:setup) and motivates domain-specific augmentation and robustness interventions (e.g., stronger brightness/contrast jitter and blur-aware training). The mean corruption accuracy (mCA) is extbf80.95%.

4.5. Efficiency and qualitative analysis

Grad-CAM qualitative localization. Fig. 10 shows Grad-CAM overlays highlighting lesion clusters and margins across representative test images. Saliency aligns with agronomic symptom descriptions and complements rule-aware rationales by indicating where the network focused when producing calibrated predictions.

Latency and memory footprint. Batch-1 latency and memory are modest. Fig. 11 shows the latency distribution; Table 14 reports summary statistics.

Table 14
Inference efficiency (batch = 1, post-warmup).

Statistic	Value	Notes
Median latency (ms/image)	**4.57**	Real-time on commodity GPU
Mean latency (ms/image)	4.51	
95th percentile (ms/image)	4.93	
Peak GPU memory	≈395 MB	From runtime logs

Limitations and future directions. While the proposed pipeline improves transparency and enables practical triage, several limitations remain.

1. **Generalization beyond the benchmark.** Even with de-duplication, *PaddyDoctor* is a single public corpus; additional evaluation across seasons, devices, and geographic regions is needed to quantify deployment shift.
2. **Robustness gaps under photometric/blur shift.** The corruption diagnosis identifies brightness and blur as critical failure modes; production deployment would require domain-specific augmentation and/or robustness training [17,23].
3. **Calibration trade-offs.** Temperature scaling improves proper scoring rules but can slightly worsen ECE on the test split; more expressive class-wise calibration (vector/Dirichlet) is a promising direction when sufficient calibration data are available [15].
4. **Validator brittleness.** The current symptom cues are simple, threshold-based detectors and can inherit sensitivity to illumination changes; learning symptom attributes (or integrating rule constraints more tightly) is a key next step.

Future work will focus on (i) stronger field-aware augmentation and domain adaptation, (ii) calibration strategies validated under deployment shift, and (iii) learned symptom detectors that better bridge visual evidence to symbolic rules for more reliable auditing.

5. Conclusion

We presented a diagnostic analysis of an attention-augmented CNN for rice disease diagnosis, coupled with post-hoc probability calibration and a post-hoc rule-aware validator. On a rigorously de-duplicated

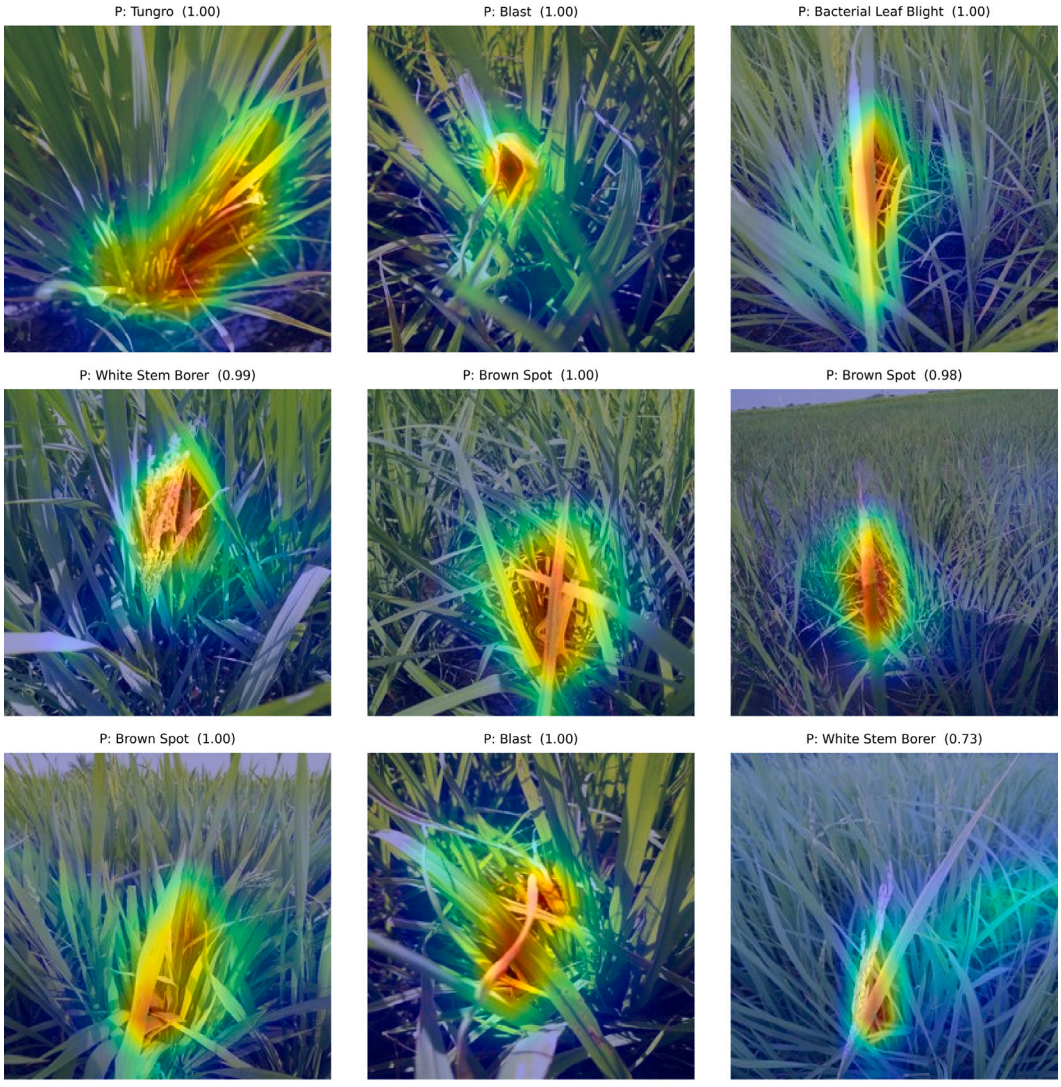


Fig. 10. Grad-CAM overlays (test samples). Titles show predicted class and calibrated confidence.

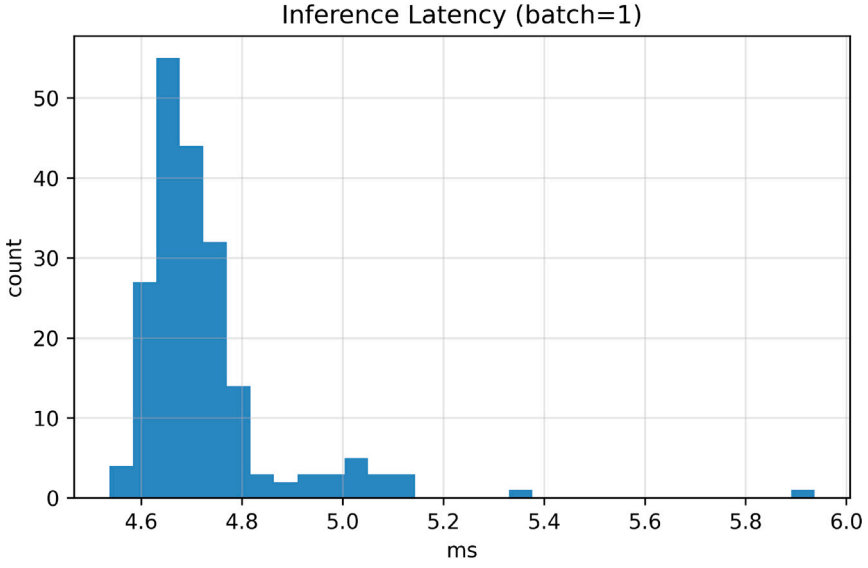


Fig. 11. Inference latency histogram (batch = 1).

held-out test set, the model attains 95.13% accuracy (weighted F1 95.14%), establishing a new, challenging baseline for this public dataset. The analysis surfaced three key gaps: (i) calibration method choice trades ECE against NLL/Brier [12,20], (ii) large performance drops under brightness/blur highlight the need for domain-specific augmentation [17,23], and (iii) the symbolic validator is post-hoc and brittle to photometric shifts [6,7]. We release artifacts to catalyze rigorous, reproducible comparisons in agricultural AI.

By pairing calibrated neural predictions with explicit agronomic rules, the system supports safer field deployment: extension workers can set confidence thresholds, auto-escalate uncertain or rule-inconsistent cases, and audit decisions. This reduces misdiagnosis risk, improves input stewardship, and strengthens traceability.

CRediT authorship contribution statement

Chatter Singh: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Amar Singh:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Conceptualization. **Sahraoui Dhelim:** Visualization, Validation, Resources, Data curation, Conceptualization.

Data and code availability

This paper utilizes the publicly available Paddy Doctor dataset [16].

Ethics statement

This article does not contain any studies with human participants or animals performed by the authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used AI-assisted tools for language polishing and consistency checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the contributors to the Paddy Doctor dataset and acknowledge the computational resources provided by their respective institutions.

Data availability

This study uses the publicly available PaddyDoctor dataset. All dataset images and labels can be accessed from the public repository provided by the original authors of PaddyDoctor, subject to their terms of use.

All split indices, preprocessing settings, and evaluation artifacts (including confusion matrices, ROC/PR curves, reliability diagrams, corruption curves, and latency logs) referenced in this paper are stored with the authors' code and experiment bundle to enable exact reproducibility.

The code and artifacts used to train, calibrate, and evaluate the proposed model will be made available by the authors upon reasonable request and/or via a public repository.

References

[1] Mohanty SP, Hughes DP, Salathé M. Using deep learning for image-based plant disease detection. Front Plant Sci 2016;7:1419. http://dx.doi.org/10.3389/fpls.2016.01419.

[2] Ferentinos KP. Deep learning models for plant disease detection and diagnosis. Comput Electron Agric 2018;145:311–8. http://dx.doi.org/10.1016/j.compag.2018.01.009.

[3] Barbedo JGA. Plant disease identification from individual lesions and spots using deep learning. Biosyst Eng 2019;180:96–107. http://dx.doi.org/10.1016/j.biosystemseng.2019.02.002.

[4] Uddin M, Alam S, Hossain M. Explainable deep learning for plant disease recognition: A framework and empirical study. Expert Syst Appl 2022;201:117063. http://dx.doi.org/10.1016/j.eswa.2022.117063.

[5] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. ICCV. 2017, http://dx.doi.org/10.1109/ICCV.2017.74.

[6] Sarker IH, Kayes A. Knowledge graphs in agriculture: A survey. Expert Syst Appl 2021;184:115578. http://dx.doi.org/10.1016/j.eswa.2021.115578.

[7] Xiong J, Sun Q, Li Y. Ontology-based agricultural knowledge representation and reasoning for disease diagnosis. Appl Soft Comput 2022;123:108935. http://dx.doi.org/10.1016/j.asoc.2022.108935.

[8] d'Avila Garcez A, Lamb LC. Neurosymbolic AI: The 3rd wave. Artificial Intelligence 2021;301:103535. http://dx.doi.org/10.1016/j.artint.2021.103535.

[9] d'Avila Garcez A, Lamb LC, Gabbay D. Neural-Symbolic cognitive reasoning. Springer; 2009.

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. CVPR. 2016, http://dx.doi.org/10.1109/CVPR.2016.90.

[11] Woo S, Park J, Lee J-Y, Kweon IS. CBAM: Convolutional block attention module. In: Proc. ECCV. 2018, http://dx.doi.org/10.1007/978-3-030-01234-2_1.

[12] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proc. ICML. 2017.

[13] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. Mach Learn 2001;45:115–35. http://dx.doi.org/10.1023/A:1010920819831.

[14] Naeini MP, Cooper G, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. Mach Learn 2017;106:1561–86. http://dx.doi.org/10.1007/s10994-017-5642-8.

[15] Kull M, Filho T Silva, Flach P. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. Pattern Recognit 2019;100:107125. http://dx.doi.org/10.1016/j.patcog.2019.107125.

[16] Petchiammal A, Kiruba SB, Murugan D, Pandarasamy A. Paddy doctor: A visual image dataset for automated paddy disease classification and benchmarking. In: Proc. 6th joint int. conf. data science & management of data. 2023, p. 1–5. http://dx.doi.org/10.1145/3570991.3570994.

[17] Sharma P, Kaur A, Garg N. Deep learning for crop disease identification in the wild: A comprehensive review. Comput Electron Agric 2022;193:106694. http://dx.doi.org/10.1016/j.compag.2021.106694.

[18] Rahman MM, Mahmud M, Shuvo P. Deep learning-based rice disease recognition: State of the art and challenges. IEEE Access 2021;9:158777–97. http://dx.doi.org/10.1109/ACCESS.2021.3130236.

[19] Lu J, Hu X, Zhuang Y, et al. A survey of deep learning for rice disease recognition in field images. Front Plant Sci 2021;12:701038. http://dx.doi.org/10.3389/fpls.2021.701038.

[20] Mathews T, Woodruff DP. On reliability diagrams: Fundamentals and improvements. Electron J Stat 2023;17(2):2470–508. http://dx.doi.org/10.1214/23-EJS2112.

[21] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. In: Proc. ICLR. 2019.

[22] Barron-Gafford J, et al. Recent advances in crop disease detection using UAV and deep learning: A review. Remote Sens 2023;15(9):2450. http://dx.doi.org/10.3390/rs15092450.
[23] Wang X, Zhang H, Zhao Y. Robust plant disease recognition with test-time augmentation and confidence estimation. Comput Electron Agric 2021;187:106285. http://dx.doi.org/10.1016/j.compag.2021.106285.
[24] Sarkar A, Singh S, Sood N. Explainable deep models for field-condition crop disease classification. Remote Sens 2022;14(22):5895. http://dx.doi.org/10.3390/rs14225895.
[25] Ghosal S, Blystone A, Singh A, et al. An explainable deep machine vision framework for plant stress phenotyping. Plant Methods 2018;14:118. http://dx.doi.org/10.1186/s13007-018-0383-6.
[26] Wang Y, Wang H, Peng Z. Rice diseases detection and classification using attention-based neural network and Bayesian optimization. Expert Syst Appl 2021;178:114770. http://dx.doi.org/10.1016/j.eswa.2021.114770.
[27] Chen J, Liu C, Zhai TF, Wang S. Detection of rice plant diseases based on deep transfer learning. J Sci Food Agric 2020;100(23):5308–16. http://dx.doi.org/10.1002/jsfa.10373.
[28] Yadav S, Kumar R, Singh SP. Deep learning-based approaches for detection and classification of rice plant diseases: A comprehensive review. Cogn Comput 2021;13:1401–23. http://dx.doi.org/10.1007/s12559-021-09806-w.
[29] Picon A, A. Alvarez-Gila M Seitz, et al. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. Plant Methods 2019;15:118. http://dx.doi.org/10.1186/s13007-019-0479-6.
[30] Wang G, Sun Y, Wang J. Automatic image-based plant disease severity estimation under field conditions. Sci Rep 2020;10:7610. http://dx.doi.org/10.1038/s41598-020-64790-6.
[31] Roy PS, Kukreja V. Vision transformers for rice leaf disease detection and severity estimation: a precision agriculture approach. J the Saudi Soc Agric Sci 2025. http://dx.doi.org/10.1007/s44447-025-00007-w.
[32] Huang X, Xu D, Chen Y, Zhang Q, Feng P, Ma Y, Dong Q, Yu F. Econv-ViT: A strongly generalized apple leaf disease classification model based on the fusion of ConvNeXt and transformer. Inf Process Agric 2025. http://dx.doi.org/10.1016/j.inpa.2025.03.001.
[33] Chen J, Liu T, Li H. Hybrid attention networks for rice disease identification in complex field images. Sensors 2021;21(20):6826. http://dx.doi.org/10.3390/s21206826.
[34] Zhang Y, Song D, Wang X. Lightweight convolutional networks for rice leaf disease identification under field conditions. Remote Sens 2022;14(3):489. http://dx.doi.org/10.3390/rs14030489.
[35] Khan S, Khan MA, Damaševičius R. A survey on rice leaf disease detection using deep learning. Agronomy 2023;13(1):30. http://dx.doi.org/10.3390/agronomy13010030.
[36] d'Avila Garcez A, Lamb LC, Gabbay D. Neuro-symbolic artificial intelligence. AI Commun 2021;34(1):1–12. http://dx.doi.org/10.3233/AIC-210084.
[37] Sarker S, Hitzler P. Neuro-symbolic AI: A review and outlook. Artif Intell Rev 2025. http://dx.doi.org/10.1007/s10462-025-11234-6, (early access).
[38] Stamper R, d'Avila Garcez A, Lamb LC. Neural-symbolic computing: An effective methodology for principled integration of learning and reasoning. Philosophical Trans Royal Society A 2020;378:20190368. http://dx.doi.org/10.1098/rsta.2019.0368.

Chatter Singh was born in Hoshiarpur, India, on December 29, 1982. He received his MCA from Maharishi Dayanand University, Rohtak, India, in 2008. He became an IEEE Student Member (S) in 2024. Mr. Singh is currently a Research Scholar in the School of Computer Applications at Lovely Professional University, Phagwara, Punjab, India. His interests include machine learning and agricultural AI for efficient, edge-based disease diagnostics. He has been a member of the International Association of Engineers, Hong Kong, China, since June 17, 2021.

Amar Singh: Prof. (Dr.) Amar Singh is a Professor at the School of Computer Application, Lovely Professional University (LPU), Phagwara, Punjab, India, with over 18 years of teaching and research experience. He holds a Ph.D. in Computer Science and Engineering from IKG Punjab Technical University and an M.Tech. in Information Technology from Maharishi Markandeshwar University. His research interests include Soft Computing, Machine Learning, and Computer Vision. Dr. Singh has published 80 research articles, authored six book chapters, and holds 24 patents. He is a member of IEEE and a life member of ISCA, and serves as an editor for the International Journal of Innovation in Multidisciplinary Scientific Research (IJIMSR).

Sahraoui Dhelim: He is an Assistant Professor in the School of Computing at Dublin City University. Previously, he served as a Senior Postdoctoral Researcher at University College Dublin from 2021 to 2024 and was a Visiting Researcher at Ulster University from 2020 to 2021. He earned his Ph.D. in Computer Science and Technology from the University of Science and Technology in Beijing, China, in 2020. He also holds a Master's degree in Networking and Distributed Systems from the University of Laghouat, Algeria, obtained in 2014, and a Bachelor's degree in Computer Science from the University of Djelfa, Algeria, obtained in 2012. He serves as an editor in the AI and Autonomous Systems Journal and as a guest editor for several reputable journals, including Electronics and Applied Sciences. His research interests encompass Data Analytics, Machine Learning, Deep Learning, IoT, and Intelligent Transportation Systems.