Fusing Perceptual Vision Experts with Multimodal Large Language Models for Explainable Plant Disease Diagnosis: From Benchmark Imagery to Real-World Robotic Field Validation⋆

Ranjan Sapkota[a,*], Konstantinos I. Roumeliotis[b,c], Pengyao Xie[a], Nikolaos D. Tselikas[b], Lirong Xiang[a] and Manoj Karkee[a,*]

[a]Cornell University, Department of Biological and Environmental Engineering, Ithaca, 14850, NY, USA
[b]University of the Peloponnese, Department of Informatics and Telecommunications, Tripoli, 22131, Greece
[c]Agricultural University of Athens, Department of Agribusiness and Supply Chain Management, Athens, 11855, Greece

ARTICLE INFO

Keywords:
Information fusion
Plant disease detection
Multimodal large language models
Explainable AI
Business decision support
Decision support systems
ConvNeXt
EfficientNet
PlantDoc
Real-world field validation
Hybrid hierarchical multi-agent framework
Conflict resolution

ABSTRACT

Accurate plant disease diagnosis in the field requires more than a class label: it requires fusing uncertain, sometimes conflicting perceptual evidence into a single, trustworthy, and actionable decision. This paper presents a **Hybrid Hierarchical Multi-Agent Framework** (H^2MAF) that performs this fusion in two stages: *decision-level fusion* of two architecturally distinct Convolutional Neural Network (CNN) perceptual experts, EfficientNet-B3 and ConvNeXt-Tiny, followed by *semantic-level fusion* in which small, open-weight Multimodal Large Language Models (MLLMs): Google Gemma 4 E4B and Alibaba Qwen3.5 4B arbitrate between the CNN experts, ground their reasoning in a structured JSON evidence artefact, and translate the fused diagnosis into a business-oriented Explainable AI (XAI) report specifying risk level, treatment urgency, and estimated financial exposure. Uniquely, we validate this fusion architecture across **three independent, increasingly realistic datasets totalling 14,364 images and 1,370 held-out test images**: (i) PlantDoc, the standard internet-scraped "in-the-wild" benchmark (2,922 images, 27 classes); and (ii, ii, and iii) two previously customized, continuously-captured field datasets provided by the Automation and Robotics Laboratory (AR Lab), Department of Biological and Environmental Engineering, Cornell University, a 20 GB acquisition campaign ("Stage 2", 4,215 images) and a 40 GB scaled follow-up campaign ("Stage 4", 7,227 images), both covering three economically important tomato/potato foliar diseases (Early Blight, Late Blight, Septoria Leaf Spot) recorded by an autonomous field-imaging platform under uncontrolled outdoor conditions. To our knowledge, this is the first study to evaluate an MLLM-based conflict-arbitration framework on real, non-public, robot-acquired agricultural imagery rather than curated or internet-sourced benchmarks alone. On PlantDoc, the MLLM fusion layer improves top-1 accuracy from 63.9% (best CNN, ConvNeXt-Tiny) to 68.5% (Gemma 4 E4B), with the gain concentrated on the 41.7% of images where the two CNN experts disagree (+7.6 points over the stronger CNN). On the Cornell Stage 2 and Stage 4 real-field datasets, both CNNs already exceed 96–99.8% accuracy, and the MLLM layer contributes smaller but still positive gains (up to 99.3% and 98.9% top-1 accuracy respectively), while vision-expert disagreement collapses to 1.7-4.1% of images empirically confirming that *MLLM arbitration utility is proportional to the rate of perceptual conflict*, not a fixed additive bonus. A novel *risk-personality calibration analysis*, replicated across all three datasets, shows that Gemma's assigned Critical-risk rate tracks the true prevalence of the urgent disease class almost exactly (0.14-0.5 percentage-point error on the two Cornell datasets), whereas Qwen systematically over-flags by 3.5-14.4 points, revealing a consistent, quantifiable "alarmist" bias that is independent of raw classification accuracy and has direct implications for deployment-time risk calibration. These findings establish both the promise and the boundary conditions of MLLM-based decision fusion for agricultural AI: substantial value under genuine perceptual uncertainty, fidelity to strong CNN consensus, and measurable, model-specific risk-assessment biases that must be calibrated before field deployment. Github Repo: Link

⋆This work is supported by the National Science Foundation (NSF) and the United States Department of Agriculture (USDA), National Institute of Food and Agriculture (NIFA), through the "Artificial Intelligence (AI) Institute for Agriculture" program under Award Numbers AWD003473 and AWD004595, and USDA-NIFA Accession Number 1029004 for the project titled "Robotic Blossom Thinning with Soft Manipulators." Additional support was provided through USDA-NIFA Grant Number 2024-67022-41788, Accession Number 1031712, under the project "Expanding UCF AI Research To Novel Agricultural Engineering Applications (PARTNER)."

*Corresponding authors
rs2672@cornell.edu (R. Sapkota); k.roumeliotis@uop.gr (K.I. Roumeliotis); px62@cornell.edu (P. Xie); ntsel@uop.gr (N.D. Tselikas); lxiang@cornell.edu (L. Xiang)
ORCID(s): 0000-0002-5417-6744 (R. Sapkota); 0000-0002-8098-1616 (K.I. Roumeliotis); add (P. Xie); 0000-0001-5799-3558 (N.D. Tselikas); 0000-0003-1573-0906 (L. Xiang); 0000-0001-5337-4848 (M. Karkee)

1. Introduction

Plant diseases account for 20–40% of global crop yield losses annually, posing a critical threat to food security and rural economies [34]. Early, accurate diagnosis is therefore essential, yet field deployment of disease detection systems remains constrained by three persistent gaps that the computer-vision and information-fusion literature has not fully bridged.

Gap 1: Single-model perceptual fragility. Benchmark datasets such as PlantVillage [15, 21] were captured under controlled laboratory conditions uniform backgrounds, consistent lighting, single leaves yielding top-1 accuracies

exceeding 99% for standard CNNs [10]. However, Barbedo [4] and Singh et al. [30] demonstrated that the same models suffer drastic performance drops when applied to real-field imagery, where cluttered backgrounds, multiple overlapping leaves, and variable illumination prevail. Any single classifier, however well trained, will therefore produce meaningfully different confidence distributions than an architecturally distinct classifier on the same challenging image, and neither is reliably "more correct" a priori: a principled way of *fusing* their evidence is required rather than trusting either signal in isolation.

Gap 2: The classification-to-action disconnect. Even a perfectly accurate classifier only outputs a class label. A farm manager responsible for a €50,000 tomato crop needs to know *how severe* the outbreak is, *which treatment* to apply, *how urgently*, and *at what financial risk* if no action is taken. Conventional pipelines offer no mechanism for this translation, leaving domain experts to interpret raw probability scores without decision support.

Gap 3: The benchmark-to-real-world validation gap. The overwhelming majority of plant disease AI studies—including our own prior work [26]—are validated exclusively on public, internet-curated, or laboratory-acquired benchmarks (PlantVillage, PlantDoc). Such datasets, however carefully constructed, are static snapshots assembled by web scraping or single-session photography; they cannot fully capture the continuous, temporally correlated, sensor-noise-laden imagery produced by an actual autonomous field-monitoring platform operating on a working farm. Recent work has confirmed that the controlled-to-field reliability gap persists even after standard mitigation techniques are applied [35], yet almost no published study has evaluated a plant disease decision-support pipeline on *genuine, non-public, robot-acquired* field data, because such data are rarely available outside the laboratories that collect them.

Our contribution. This paper proposes a **Hybrid Hierarchical Multi-Agent Framework** (H^2MAF) that closes all three gaps through a principled two-stage information-fusion architecture and, critically validates that architecture on three progressively more realistic datasets—including two previously unpublished, closed, field-acquired datasets collected by the authors. The framework operates in three phases, implementing decision-level fusion followed by semantic-level fusion:

1. **Perceptual Classification (decision-level fusion input).** Two architecturally distinct CNNs—EfficientNet-B3 [31] and ConvNeXt-Tiny [19]—are independently fine-tuned on each dataset, producing probability distributions over the target disease classes.
2. **Contextual Alignment.** CNN outputs are serialised alongside crop- or disease-specific business context (crop value, pathogen identity, weather conditions, risk-aversion level) into a structured JSON artefact that grounds MLLM reasoning and suppresses hallucination.
3. **Cognitive Reasoning (semantic-level fusion).** Two small Multimodal LLMs, Google Gemma 4 E4B and Alibaba Qwen3.5 4B, receive the leaf image plus the JSON artifact and produce a structured Explainable AI (XAI) report containing a final diagnosis, arbitration reasoning, visual symptom description, risk level, business recommendation, and treatment window.

This separation exploits the complementary strengths of each component: CNNs excel at fine-grained texture discrimination; MLLMs contribute commonsense agricultural knowledge, semantic conflict resolution, and natural-language explanation generation—tasks for which they are inherently suited. The framework is *model-agnostic* and *dataset-agnostic*: any fine-tuned CNN or instruction-following MLLM can be plugged in, and the same three-phase pipeline is applied, without architectural changes, to three datasets that differ radically in acquisition modality, class count, and visual difficulty.

The main contributions of this work are:

- The first systematic study of MLLM-driven conflict arbitration between competing CNN perceptual experts validated not only on an internet-curated "in-the-wild" benchmark (PlantDoc) but also on two previously unpublished, continuously-captured, robot-acquired real-world field datasets (Cornell Stage 2, 20 GB; Cornell Stage 4, 40 GB), spanning 14,364 images and 1,370 held-out test images in total.
- Empirical demonstration, replicated across all three datasets, that MLLM arbitration value is a function of the vision-expert disagreement rate: substantial gains (+7.6 points) at a 41.7% conflict rate (PlantDoc), and comparatively marginal gains at conflict rates of 1.7–4.1% (Cornell Stage 2/4), where the CNN experts already agree on 95.9–98.3% of images.
- A rigorous analysis of MLLM override behaviour across all three datasets, showing that models which override *both* CNNs independently achieve only 0–17.6% accuracy, establishing a design principle against unconstrained MLLM autonomy that generalises beyond a single dataset.
- A novel *risk-personality calibration* analysis quantifying how closely each MLLM's assigned Critical-risk rate tracks the true prevalence of the most urgent disease class, replicated across the two Cornell datasets, revealing that Gemma is consistently well-calibrated (0.14–0.5-point error) while Qwen is consistently alarmist (3.5–14.4-point over-flagging).
- A fully reproducible, open-source pipeline integrating CNN training, JSON artefact generation, MLLM inference, and automated evaluation, applied uniformly across three independent datasets of increasing scale and realism.

The remainder of this paper is structured as follows. Section 2 reviews the related literature on plant disease classification, multimodal LLMs, XAI, and information fusion

for decision support. Section 3 describes all three datasets: PlantDoc and the two Cornell real-world field datasets: and the experimental train/validation/test protocol for each. Section 4 presents the three-phase H^2MAF architecture in detail. Section 5 defines the experimental setup and evaluation metrics. Section 6 reports quantitative results on each of the three datasets individually, followed by a cross-dataset synthesis. Section 7 interprets the findings, discusses limitations, and identifies future directions. Section 8 concludes the paper.

2. Related Work

2.1. Plant Disease Classification

Deep learning has dominated plant disease classification since Mohanty et al. [21] demonstrated 99% accuracy on PlantVillage using AlexNet. Subsequent studies systematically benchmarked ResNet [13], DenseNet [14], and EfficientNet [31] variants, consistently achieving near-perfect accuracy on controlled images while confirming large performance drops on real-field data [4, 16]. Singh et al. [30] introduced PlantDoc to address this gap; since its release, it has become the canonical benchmark for "in-the-wild" classification, with reported accuracies typically in the 60–75% range for single CNNs.

ConvNeXt [19] represents a post-ViT modernisation of the convolutional paradigm, matching Vision Transformer [8] performance while retaining CNN inductive biases—crucial for small-data regimes. We adopt ConvNeXt-Tiny as our second perceptual expert precisely because it outperforms Transformer-based alternatives on small training sets such as PlantDoc's 2,269-image training split.

2.2. Multimodal LLMs in Visual Understanding

The emergence of instruction-following Multimodal LLMs (MLLMs)—from Flamingo [2] and GPT-4V [1] to open-weight alternatives such as Gemma 4 [11] and Qwen3 [23]—has introduced a new paradigm for visual reasoning. Unlike CNNs, which learn discriminative features through supervised fine-tuning, MLLMs bring pre-trained commonsense, domain knowledge, and language generation capabilities that enable rich, context-aware explanations [36].

Roumeliotis et al. [26] compared GPT-4 fine-tuned for plant disease classification against ResNet-50, finding that LLMs can encode domain knowledge competitive with supervised CNNs. However, direct MLLM classification of fine-grained disease textures without any CNN assistance remains unreliable at the confidence levels required for actionable recommendations. Our work differs fundamentally by *not* replacing CNNs with MLLMs but instead positioning MLLMs as arbitrators that consume pre-computed CNN signals—a form of hierarchical decision fusion in which each modality contributes at the layer where it is strongest. We also note that none of these prior studies–including our own–were validated beyond internet-scraped or laboratory-curated imagery; the present work is, to our knowledge, the first to test this fusion paradigm against genuine, closed, robot-acquired field data.

2.3. Explainable AI in Agriculture

XAI methods such as LIME [24], SHAP [20], and Grad-CAM [27] have been applied to visualise CNN decisions in plant pathology [5], but they produce pixel-level saliency maps that require agronomist expertise to interpret. No existing work translates CNN classifications into structured, actionable crop management recommendations with quantified financial impact—the primary novelty of our XAI layer.

2.4. Transfer Learning and Domain Adaptation

Transfer learning from ImageNet pre-trained weights has become the de facto training protocol for plant disease classification [16]. Mohanty et al. [21] demonstrated that even shallow fine-tuning of VGG-16 and AlexNet on PlantVillage yields 97–99% accuracy, establishing the paradigm. Subsequent work explored fine-tuning depth: Ferentinos [10] compared VGG-16, AlexNet, GoogLeNet, and a custom architecture, finding that deeper unfreeze strategies and higher-resolution inputs consistently improve generalisation on in-the-wild samples.

However, transfer from ImageNet to agricultural images presents a domain gap: ImageNet features are tuned for object recognition (edges, shapes, object parts), while disease detection requires sensitivity to subtle textural patterns (concentric ring formation, sporulation density, lesion colouration gradients) that are not well-represented in ImageNet classes. Partial-unfreeze strategies [31] balance preservation of low-level texture detectors with task-specific adaptation of mid- and high-level features. In the present work, we apply partial unfreeze (top 30%) for EfficientNet-B3 and full fine-tune for ConvNeXt-Tiny [19] uniformly across all three datasets, the latter being justified by ConvNeXt's stronger per-layer feature reuse from its inverted bottleneck design.

2.5. Ensemble Methods and Multi-Expert Model Fusion

Ensembling multiple models to improve robustness has a long history in computer vision, from classical bagging and boosting to modern deep ensemble and mixture-of-experts architectures. In the plant disease domain, ensemble approaches combining CNNs with different architectures or training regimes consistently outperform single-model baselines by 3–8 percentage points on challenging benchmarks [16]. However, standard ensembles aggregate predictions through fixed rules (majority vote, average probability) without any semantic understanding of why the models disagree or what the disagreement implies operationally. More recently, Roumeliotis et al. [25] proposed a modular agentic AI framework with trust-aware orchestration and retrieval-augmented reasoning for visual classification, demonstrating that orchestrator-agent trust mechanisms can improve calibration and reduce overconfident misclassifications—a direction complementary to the semantic arbitration approach adopted in this work.

Our framework differs from traditional ensembling in a fundamental way: the second-stage arbitration is performed by a model (MLLM) that has not been trained

on the target dataset and has no direct access to class probability statistics. The MLLM introduces a new, orthogonal information source—pre-trained visual-semantic knowledge—that is architecturally decoupled from the first-stage CNN uncertainty. This is closer in spirit to a human expert reviewing conflicting laboratory reports: the expert does not simply average the reports, but applies domain knowledge to determine which is more credible given the available visual evidence. This work extends that principle by testing whether the same fusion logic transfers unchanged from a 27-class, internet-curated benchmark to a 3-class, continuously-captured, real-world robotic field dataset—a substantially different information-fusion regime in which the CNN experts rarely disagree at all.

2.6. Prompt Engineering for Structured Output

Instructing LLMs to produce structured output (JSON, XML, SQL) is an active research area. Naive prompting frequently results in verbose free-form responses that are difficult to parse reliably. Key techniques include explicit format specification [1], chain-of-thought suppression for latency-sensitive applications [23], and constrained decoding approaches that enforce syntactic validity.

Across all three datasets used in this study, we found a consistent distinction between the two MLLM families: Gemma 4 E4B follows JSON formatting instructions reliably in zero-shot mode, while Qwen3.5 4B's built-in chain-of-thought reasoning mode consumes the generation budget with verbose reasoning before producing structured output. Suppressing thinking mode via `enable_thinking=False` in the chat template resolved this issue, yielding coverage above 96.8% on every dataset. This finding has broader implications for deploying reasoning-capable MLLMs in production pipelines where structured output is mandatory.

Power [22] surveyed the evolution of Decision Support Systems (DSS) from rule-based expert systems to data-driven models. Wolfert et al. [34] identified smart-farming data integration as a key challenge for modern agricultural DSS. Our framework takes a first step toward addressing this challenge by embedding crop- and disease-specific economic context—including estimated crop value, pathogen identity, and a risk-aversion level—into the structured JSON artefact consumed by the MLLM, enabling the reasoning layer to generate outputs that go beyond bare classification labels to include treatment urgency, risk level, and estimated financial impact, consistently across benchmark and real-world data.

3. Datasets: From Internet Benchmark to Real-World Robotic Field Validation

A central methodological contribution of this study is the validation of the same H^2MAF pipeline on three datasets of markedly different provenance, scale, and realism. Table 1 summarises the three datasets; Fig. 1 provides representative sample images from each; the remainder of this section describes each in turn.

3.1. PlantDoc: An Internet-Curated "In-the-Wild" Benchmark

PlantDoc [30] is the primary public benchmark for "in-the-wild" plant disease classification. Unlike PlantVillage [15], whose images were captured under controlled laboratory conditions, PlantDoc consists of images scraped from the internet, reflecting the visual diversity of casual field photography: cluttered multi-leaf backgrounds, variable illumination, mixed healthy/diseased content within a single frame, and significant compression artefacts.

The Kaggle Version 7 of PlantDoc [28] used in this study comprises **2,922 images** organised into a predefined train/test split: 2,670 training images across 27 class folders and 252 test images across 26 class folders (one class, `Tomato_two_spotted_spider_mites_leaf`, has only 2 training images and is absent from the test set, and is therefore excluded from all experiments). The final dataset used for training and evaluation covers **27 classes** spanning 13 plant species, including both diseased and healthy leaf categories. Table 2 summarises the per-class distribution; significant imbalance is present (`Tomato_leaf_yellow_virus`: 223 train images vs. `Bell_pepper_leaf`: 34, a 6.6× ratio), necessitating Focal Loss [18] during CNN training.

A stratified 85/15 split of the 2,670 training images creates a validation set for model selection: 2,269 images for training and 401 for validation. The 252 test images from the original PlantDoc split are held out entirely and used only for final evaluation.

3.2. Cornell Real-World Field Datasets: Stage 2 and Stage 4

To address the benchmark-to-real-world validation gap identified in Section 1, this study incorporates two customized, closed datasets acquired by co-authors at Cornell University. Both datasets target three economically important foliar diseases of tomato and potato: **Early Blight** (*Alternaria solani*), **Late Blight** (*Phytophthora infestans*), and **Septoria Leaf Spot** (*Septoria lycopersici*).

3.2.1. Acquisition Protocol

Unlike PlantDoc, whose images are independently sourced internet photographs, the Cornell datasets consist of continuous field-image captures acquired using a customized phenotyping platform adapted from the Amiga robot. Field trials were conducted at the Mountain Research Station (MRS) in Waynesville, North Carolina, and the Mountain Horticultural Crops Research and Extension Center (MHCREC) in Mills River, North Carolina, during the 2025 growing season. The experimental plots included multiple tomato cultivars and were subjected to controlled disease inoculation to facilitate the development of naturally progressing foliar symptoms.

The phenotyping platform was equipped with five independent imaging units, each consisting of a stereo camera and a custom active strobe-lighting array. As the robot navigated along the crop rows, plant canopies were imaged from both sides. Active illumination was used during image

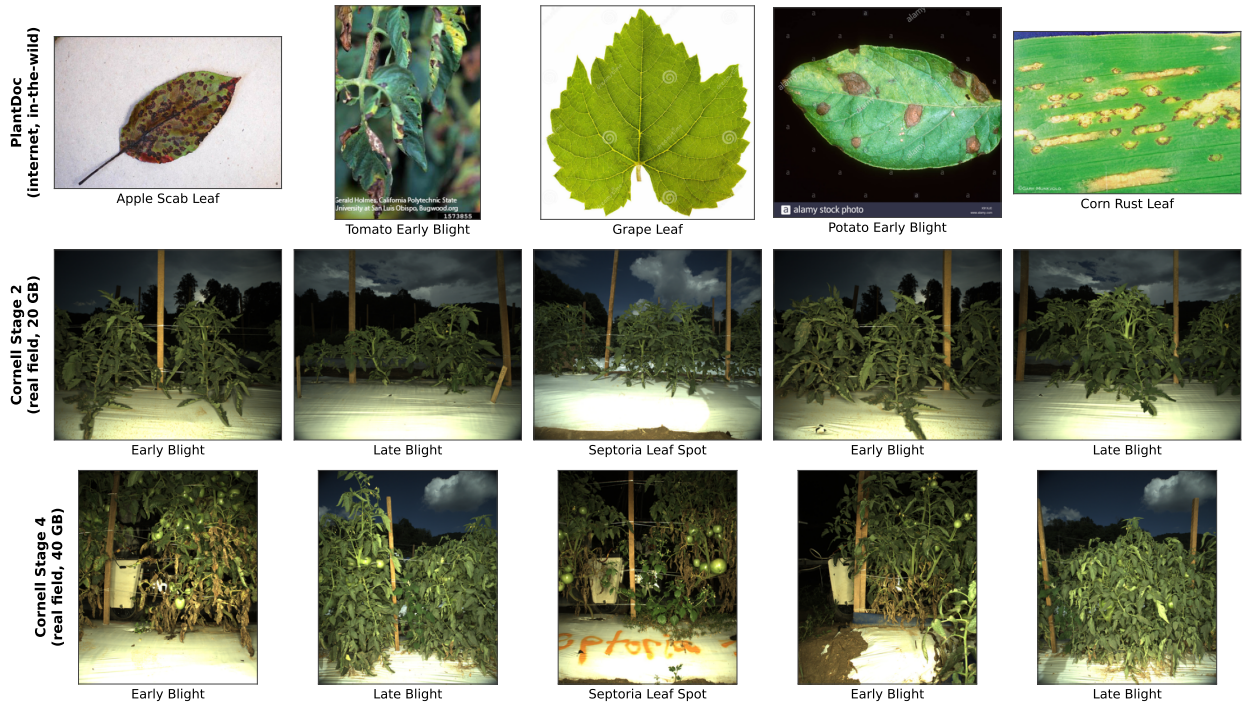

Figure 1: Representative sample images from the three datasets used in this study. Top row: PlantDoc, internet-scraped "in-the-wild" photographs (27 classes). Middle and bottom rows: Cornell Stage 2 and Stage 4, continuously-captured, autonomous field-robot imagery (3 classes). The visual contrast between crowd-sourced internet photography and continuous field-video frames illustrates the acquisition-modality gap discussed throughout this section.

Table 1
Overview of the three datasets used in this study. "Acquisition" describes how images were captured; "Setting" indicates whether the source is a public internet-curated benchmark or a closed, real-world field dataset shared under a research collaboration.

Dataset	Images	Classes	Test N	Acquisition	Setting
PlantDoc [30, 28]	2,922	27	252	Internet-scraped photographs	Public, in-the-wild benchmark
Cornell Stage 2 (this work)	4,215	3	403	Autonomous field-robot video frames, 20 GB	Closed, real-world field data
Cornell Stage 4 (this work)	7,227	3	715	Autonomous field-robot video frames, 40 GB	Closed, real-world field data
Total	**14,364**	—	**1,370**	—	—

acquisition to reduce image-distribution shifts caused by variations in ambient field lighting. Data were collected at four temporal stages between July and August 2025 to capture disease progression, with Stage 2 acquired during July 16–18 and Stage 4 during August 15–17.

Per-image timestamps embedded in the raw data (accurate to the millisecond) confirm that frames within each class folder were captured in rapid, near-continuous succession (typically 0.3–1 s intervals). Stage 4 therefore represents an independently acquired, later-season and larger follow-up campaign rather than a strict superset of Stage 2. This acquisition modality is fundamentally different from, and considerably closer to real deployment conditions than, the single-frame internet photography underlying PlantDoc: it reflects the imagery produced by an autonomous field phenotyping platform operating under realistic outdoor conditions, including natural variation in viewing distance, angle, plant architecture, and residual environmental illumination. Prior to release, both datasets underwent a documented, multi-step expert curation pipeline performed by the AR Lab: (i) frames were manually screened to separate genuine diseased-leaf content from soil and background clutter; (ii) frames in which soil had been erroneously flagged as diseased tissue were relabelled back to the correct (healthy / background) category; (iii) blank, fully dark, or otherwise unusable frames were discarded; and (iv) only RGB imagery was retained. This expert-verified labelling pipeline provides a level of label quality assurance that is difficult to guarantee for internet-scraped datasets such as PlantDoc, whose labels derive from the original photographer's caption or uploader-supplied tag rather than a dedicated plant-pathology screening protocol.

3.2.2. Cornell Stage 2 (20 GB)

The Stage 2 release comprises **4,215 images** across the three disease classes: Early Blight (1,533), Late Blight (1,478), and Septoria Leaf Spot (1,204). No predefined train/test split accompanies the raw data, so we construct a stratified 80/10/10 train/validation/test split (random seed

Figure 2: The customized robotic platform used for collecting disease images in a tomato field.

Table 2
PlantDoc class distribution (27 classes used in experiments).

Class	Train	Test
Apple Scab Leaf	83	10
Apple Leaf	79	9
Apple Rust Leaf	96	10
Bell Pepper Leaf	34	8
Bell Pepper Leaf Spot	74	9
Blueberry Leaf	106	11
Cherry Leaf	47	10
Corn Gray Leaf Spot	63	4
Corn Leaf Blight	182	12
Corn Rust Leaf	107	10
Peach Leaf	103	9
Potato Early Blight	157	14
Potato Late Blight	200	8
Raspberry Leaf	112	7
Soyabean Leaf	57	8
Squash Powdery Mildew	124	6
Strawberry Leaf	88	8
Tomato Early Blight	79	9
Tomato Septoria Leaf Spot	145	12
Tomato Leaf	44	8
Tomato Bacterial Spot	101	9
Tomato Late Blight	101	10
Tomato Mosaic Virus	44	10
Tomato Yellow Virus	223	15
Tomato Mold	85	6
Grape Leaf	63	12
Grape Black Rot	71	8
Total	**2,670**	**252**

42) for reproducibility, consistent with the PlantDoc protocol's use of a held-out test partition. After the automated pipeline (Section 4) filters a small number of corrupted or unreadable frames, the evaluated test partition comprises **403 images**: 147 Early Blight, 141 Late Blight, and 115 Septoria Leaf Spot.

3.2.3. Cornell Stage 4 (40 GB)

The Stage 4 release is a substantially larger, independently acquired follow-up campaign comprising **7,227 images**: Early Blight (3,295), Late Blight (1,520), and Septoria Leaf Spot (2,412)—nearly double the total volume of Stage 2, with the largest proportional increase in the Early Blight and Septoria classes. The same stratified 80/10/10 split protocol (seed 42) yields a nominal test partition of 723 images, of which **715** are successfully processed end-to-end by the pipeline (329 Early Blight, 146 Late Blight, 240 Septoria Leaf Spot); the eight excluded frames failed automated image-integrity checks and were skipped by the data loader rather than mis-imputed.

3.2.4. Business Context for the Cornell Diseases

Consistent with the PlantDoc business-context module (Section 4), each Cornell disease class is mapped to a disease-specific business-intelligence record containing the causal pathogen, an estimated crop value at risk (€35,000 for Early and Late Blight, €50,000 for Septoria Leaf Spot, reflecting typical potato/tomato field values), a risk-aversion level, and a treatment-urgency label. Late Blight is the only class flagged *high* risk-aversion / *urgent* treatment-urgency in this lookup table, reflecting its well-documented capacity for rapid, epidemic-scale crop destruction (*Phytophthora infestans* was the causal agent of the Irish Potato Famine);

Table 3
Comparison of plant disease image datasets relevant to this study.

Dataset	Images	Classes	Setting	CNN Acc.
PlantVillage [15]	54,306	38	Lab, public	>99%
PlantDoc [30]	2,598	27	Field, public	60–75%
PlantDoc v7 (this work)	2,922	27	Field, public	63.9%
Cornell Stage 2 (this work)	4,215	3	Field, closed	99.8%
Cornell Stage 4 (this work)	7,227	3	Field, closed	98.3%

this single design choice becomes analytically important in Section 6, where it provides a ground-truth reference against which each MLLM's risk-assignment calibration can be quantitatively measured.

3.3. Cross-Dataset Comparison and Rationale

Table 3 contextualises all three datasets within the broader landscape of plant disease image datasets, and Table 1 (above) summarises their scale. PlantVillage's controlled acquisition protocol makes it unsuitable for evaluating robustness to real-world conditions; we therefore do not use it directly, but cite its reported accuracy ceiling as context. PlantDoc's substantially lower CNN accuracy (60–75% in the literature) directly reflects the domain gap between laboratory and internet-sourced field imagery, and we retain it here precisely because it provides a challenging, conflict-rich classification scenario (41.7% CNN disagreement rate; Section 6) that motivates and stress-tests the MLLM arbitration layer. The two Cornell datasets complement PlantDoc by offering something no public benchmark can: genuine, continuously-captured, non-curated field imagery from an operational research platform, at two different scales (20 GB and 40 GB), enabling us to test whether the H^{2}MAF fusion architecture—and its associated findings regarding conflict-dependent MLLM utility and risk-personality calibration—generalise beyond a single dataset and a single acquisition modality.

We emphasise an important, honest caveat regarding the Cornell datasets' near-ceiling CNN accuracy (96–99.8%; Section 6), which is far higher than PlantDoc's 59–64%. Three factors jointly explain this gap, and all are discussed transparently rather than presented as an unqualified strength: first, the Cornell task is genuinely easier in a taxonomic sense (3 visually distinct diseases on a narrow host range vs. 27 classes across 13 species with substantial inter-class visual similarity); second, because images are drawn from continuous video sequences, the stratified random split may place temporally adjacent (and hence highly similar) frames of the same physical lesion into different partitions, which likely inflates measured accuracy relative to a deployment scenario in which the system encounters a truly novel plant; and third, the Cornell images are full plant-canopy views captured at a consistent distance by the robot platform, whereas PlantDoc images are individual leaf photographs with highly variable framing, background clutter, and scale—a difference in image format that further contributes to the difficulty gap. We return to this limitation in Section 7 and recommend session-level (rather than frame-level) splitting and localised lesion-patch cropping for future work using this data modality.

4. Methodology

4.1. Framework Overview

Fig. 3 illustrates the H^{2}MAF pipeline, which is applied identically to all three datasets described in Section 3. A leaf image enters Phase 1, where two CNN perceptual experts independently produce top-K probability distributions over the dataset's target classes (27 for PlantDoc; 3 for each Cornell dataset). Phase 2 assembles these predictions alongside dataset-specific business context into a structured JSON artefact. Phase 3 routes the original image plus the JSON artefact to two MLLM reasoning agents, each producing a complete XAI report. This is a two-stage information-fusion pipeline: Phase 1 performs *decision-level fusion* implicitly available to Phase 3 (both CNN decisions are exposed side by side), and Phase 3 performs *semantic-level fusion*, using pre-trained visual-linguistic knowledge to weigh, reconcile, or override the two decision-level signals.

4.2. Phase 1: Perceptual Classification

4.2.1. EfficientNet-B3

EfficientNet-B3 [31] employs *compound scaling*—jointly increasing network depth, width, and input resolution—to achieve an efficient accuracy-parameter trade-off. Its design philosophy makes it particularly suited to fine-grained texture tasks where spatial resolution matters (e.g., spotting concentric ring patterns characteristic of early blight).

We adopt a *partial-unfreeze* transfer learning strategy: the early feature-extraction blocks (`features[0-5]`) are frozen to preserve low-level ImageNet representations, while the top three blocks (`features[6-8]`), the adaptive pool, and the classification head are unfrozen. This prevents catastrophic forgetting while allowing task-specific adaptation, and is applied identically on all three datasets.

4.2.2. ConvNeXt-Tiny

ConvNeXt-Tiny [19] modernises the classical ResNet paradigm by incorporating Vision Transformer design principles, large-kernel depthwise convolutions (7×7), inverted bottleneck structure, LayerNorm, and GELU activations, while retaining the translation invariance and locality inductive biases of standard convolutions. We adopt a *full fine-tune* strategy with stochastic depth [7] (drop rate 0.1) and RandAugment [7] augmentation, which we found critical for generalisation on all three datasets, particularly PlantDoc's small training set.

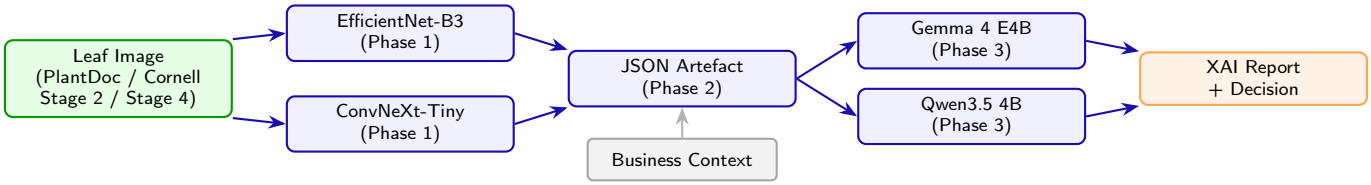

Figure 3: H²MAF pipeline, applied uniformly across the three datasets. Two CNN experts (Phase 1) feed a structured JSON artefact (Phase 2) which, together with the original image, is processed by two MLLM reasoning agents (Phase 3) to produce structured XAI reports.

Table 4
Fine-tuning hyperparameters for the two CNN perceptual experts, applied uniformly to PlantDoc and both Cornell datasets.

Parameter	EfficientNet-B3	ConvNeXt-Tiny
Input resolution	300×300	224×224
Pre-trained weights	ImageNet-1k	ImageNet-1k
Fine-tune strategy	Partial (top 30%)	Full network
Optimizer	AdamW	AdamW
Learning rate	1e-4	4e-4
LR schedule	Cosine annealing	Cosine + 5-ep warmup
LR minimum	1e-6	1e-6
Weight decay	1e-2	5e-2
Batch size (total)	256	256
GPUs	4× H100 (DDP)	4× H100 (DDP)
Epochs	50	50
Loss function	Focal Loss (γ=2)	Focal Loss (γ=2)
Augmentation	Flip, Rotate, Jitter, GaussianBlur	Flip, Rotate, Jitter, GaussianBlur, RandAugment

The contrasting design philosophies (compound scaling vs. modernised convolution) are deliberate: architecturally distinct experts produce meaningfully different confidence distributions, creating genuine conflicts that motivate MLLM arbitration on PlantDoc, even though the same experts converge to near-unanimous agreement on the visually narrower Cornell datasets (Section 6). Table 4 summarises the hyperparameters for both models, which are identical across all three datasets except for the number of output classes.

4.2.3. Training Infrastructure

All CNN training runs (six in total: two architectures × three datasets) use PyTorch DistributedDataParallel (DDP) [17] across four NVIDIA H100 NVL GPUs (CUDA 12.8, PyTorch 2.x+cu124). Focal Loss [18] with inverse-frequency class weights is used to mitigate class imbalance, which is severe for PlantDoc (up to 6.6×) and comparatively mild for the Cornell datasets (up to 2.9× between Late Blight and Early Blight in Stage 4). Best checkpoints are selected based on validation Macro F1, computed at the end of each epoch.

4.3. Phase 2: Contextual Ingestion and Alignment

A structured JSON artefact is constructed for each test image, containing: (a) the image path and ground-truth label (for evaluation only, not shown to the MLLM); (b) the top-3 predicted classes and confidence scores from each CNN expert; and (c) a *business context* object. For PlantDoc, this object is derived from a crop-to-value lookup table mapping plant species to estimated crop value (€), risk-aversion level, and current weather conditions. For the Cornell datasets, the business context is disease-specific rather than species-specific (Section 3.2), containing the causal pathogen, crop value, risk-aversion level, and treatment urgency.

```
{
  "image_id": "test_Tomato_Early_blight_leaf_3.jpg",
  "vision_experts": {
    "EfficientNet-B3": [
    {"class":"Tomato_Early_blight_leaf","confidence":0.72},
      {"class":"Tomato_leaf","confidence":0.15}
    ],
    "ConvNeXt-Tiny": [
      {"class":"Tomato_leaf","confidence":0.58},
    {"class":"Tomato_Early_blight_leaf","confidence":0.36}
    ]
  },
  "business_context": {
    "crop_value": 50000,
    "current_weather": "variable",
    "risk_aversion": "high"
  }
}
```

This explicit anchoring of MLLM reasoning to pre-computed CNN signals has two important effects: it reduces hallucination (the MLLM cannot fabricate a class that is absent from the top-3 predictions without explicitly overriding

both experts), and it provides verifiable evidence for any decision the MLLM makes.

4.3.1. Prompt Design

Careful prompt engineering is essential for reliable structured output from instruction-following MLLMs. We use a two-part prompt structure, held constant across all three datasets apart from the class-name vocabulary and business-context fields.

System prompt (Gemma 4 E4B, PlantDoc variant):

> *You are an expert agricultural AI assistant specialising in plant disease diagnosis and business intelligence for crop management. You will receive a leaf image alongside predictions from two independent computer-vision models. Your role is to synthesise these predictions, resolve any disagreement using visual evidence, and deliver a precise, actionable business recommendation.*

System prompt (Qwen3.5 4B, condensed):

> *You are a plant disease diagnosis assistant. You must respond with ONLY a JSON object. No thinking. No explanation. No markdown. Just the JSON.*

For the Cornell datasets, the equivalent system prompts explicitly enumerate the three possible diagnoses (`Early_Blight`, `Late_Blight`, `Septoria_Leaf_Spot`) and note that all three are potato/tomato foliar diseases, focusing the MLLM's semantic reasoning on the narrower, disease-specific vocabulary appropriate to that dataset.

The user prompt, identical in structure across datasets, presents: (i) the leaf image; (ii) the two CNN prediction blocks with class names and confidence scores; (iii) the business context; and (iv) a strict instruction to output only a JSON object with exactly seven fields. For Qwen3.5, the condensed system prompt was found empirically to be necessary on every dataset: the full system prompt caused the model to enter its chain-of-thought reasoning mode, exhausting the 2,048-token generation budget with verbose analysis before reaching the JSON output.

4.3.2. Output Validation and Error Recovery

MLLM responses are validated with a two-pass parser: (1) direct regex-based JSON extraction, and (2) re-attempt after stripping Markdown code fence delimiters (“‘`json` / “‘). Responses that fail both passes are recorded as parse errors; their images are excluded from accuracy computation but are counted against coverage. In production deployment, we recommend implementing a retry loop with a simplified prompt as a third fallback, which would eliminate the small residual parse failures observed on each dataset (0.8–3.2%, depending on model and dataset).

4.4. Phase 3: Multimodal Cognitive Reasoning

4.4.1. MLLM Selection

We select two small, open-weight MLLMs from distinct model families to demonstrate framework generalisability across datasets:

Google Gemma 4 E4B [11] employs Per-Layer Embeddings (PLE) for parameter efficiency, achieving 4.5B effective parameters despite 8B stored weights. It supports text, image, and audio modalities with a 128K-token context window and a configurable thinking/reasoning mode.

Alibaba Qwen3.5 4B [23] is a dense, unified vision-language model with a 4B-parameter transformer backbone and a built-in visual encoder. It supports text and image inputs with strong multimodal reasoning capabilities. During inference, `enable_thinking=False` is passed to `apply_chat_template` to suppress its chain-of-thought reasoning mode and force direct structured output.

Both models are loaded via the HuggingFace `transformers` library (`AutoModelForImageTextToText`) with `device_map="auto"` on a single H100 GPU, for all three datasets.

4.4.2. Prompt Engineering

Each MLLM receives a *system prompt* establishing its role as an agricultural AI expert, followed by a structured *user prompt* comprising: (i) the leaf image; (ii) the formatted CNN predictions with confidence scores; (iii) the business context; and (iv) an explicit instruction to respond with only a valid JSON object containing seven mandatory fields: `final_diagnosis`, `confidence`, `arbitration_reasoning`, `visual_symptoms_observed`, `risk_level`, `business_recommendation`, and `treatment_window_hours`. This seven-field schema is held fixed across all three datasets, enabling direct cross-dataset comparison of MLLM behaviour (Section 6).

5. Experimental Setup

All experiments, across all three datasets, are conducted on a compute server equipped with four NVIDIA H100 NVL GPUs (80 GB VRAM each), CUDA 12.8, and PyTorch 2.x compiled for CUDA 12.4 (cu124). CNN fine-tuning runs for 50 epochs with checkpoint selection based on validation Macro F1; wall-clock training time is approximately 2 hours per model on PlantDoc and 2–4 hours per model on the larger Cornell datasets. MLLM inference is performed on a single H100 GPU with `device_map="auto"` and `dtype="auto"` (BF16); each test image requires a single forward pass per MLLM.

Hyperparameter selection. All CNN hyperparameters listed in Table 4 were set based on established conventions for the respective architectures on small-data classification tasks and applied without dataset-specific tuning, in order to isolate the effect of dataset realism (rather than per-dataset hyperparameter optimisation) on the reported results. The partial-unfreeze depth for EfficientNet-B3 (top 30%, corresponding to `features[6-8]`) was chosen to preserve low-level texture detectors learned on ImageNet while adapting higher-level semantic representations to the disease classification task. The full fine-tune strategy for ConvNeXt-Tiny is consistent with the original ConvNeXt paper's recommendation for transfer learning on small datasets, where stochastic depth and AdamW weight decay provide sufficient regularisation to prevent overfitting.

MLLM generation parameters. Both MLLMs use `temperature=1.0`, `max_new_tokens=2048`, and `do_sample=True`, held constant across all three datasets. Temperature scaling parameters (`top_p`, `top_k`) are set to their respective model defaults to avoid interfering with the models' calibrated sampling distributions.

Evaluation metrics. We report Top-1 accuracy, Macro F1 (unweighted mean over all target classes), and Weighted F1 (weighted by class support), computed independently for each dataset. Macro F1 is the primary metric because it is invariant to class imbalance and therefore reflects performance on minority disease classes, which are of greatest agronomic significance.

MLLM output validation. MLLM responses are parsed with the two-pass JSON extractor described in Section 4. Responses that fail both passes are recorded as parse errors and excluded from accuracy computation (coverage is reported separately for each dataset and model).

Cross-dataset comparability. Because the three datasets differ in class count (27 vs. 3), test-set size (252 vs. 403 vs. 715), and label taxonomy, absolute accuracy values are not directly comparable across datasets; instead, our cross-dataset synthesis (Section 6) focuses on *relative* quantities that are meaningful regardless of class count—the accuracy gain of MLLM arbitration over the best CNN, the vision-expert agreement rate, and the calibration of MLLM risk assignment against known class prevalence.

6. Results

Results are reported per dataset (Sections 6.1–6.3) followed by a cross-dataset synthesis (Section 6.4) that constitutes the primary evidence for the paper's central claims regarding conflict-dependent MLLM utility and risk-personality calibration.

6.1. PlantDoc Results

6.1.1. Full Per-Class Classification Report

Table 5 reports per-class precision, recall, and F1 for ConvNeXt-Tiny and per-class accuracy for both MLLM agents (Gemma 4 E4B and Qwen3.5 4B) across all 27 classes. The table highlights the heterogeneity of performance: six classes achieve F1 $\geq$ 0.84 for ConvNeXt-Tiny (Strawberry, Grape Leaf, Grape Black Rot, Raspberry, Squash Powdery Mildew, Corn Rust), while six classes fall below F1 = 0.42 (Corn Gray Leaf Spot, Potato Early Blight, Potato Late Blight, Tomato Mold, Tomato Bacterial Spot, Tomato Mosaic Virus).

The six hardest classes share a common characteristic: they exhibit high visual similarity to other classes within the same genus or infection type. Potato Early Blight and Potato Late Blight, for example, both produce necrotic lesions on potato foliage; the distinguishing features (lesion shape, water-soaked border, sporulation pattern) require fine spatial resolution and semantic knowledge that the available training images (14 and 8 test samples respectively) do not sufficiently represent. Tomato Mold (*Botrytis cinerea*) and Tomato Bacterial Spot both produce small dark lesions, making textural discrimination especially difficult.

The Gemma column reveals that MLLM gains are concentrated in exactly these difficult classes: Tomato Septoria (+8.3%), Tomato Early Blight (+11.1%), Tomato Mosaic Virus (+20%), Apple Scab (+20%), and Corn Gray Leaf Spot (+25%), while structurally distinctive, easier classes (Strawberry, Grape, Corn Rust) show no change since the CNN is already near-perfect. Qwen3.5 4B shows a complementary pattern: it outperforms Gemma on Tomato Bacterial Spot (44% vs. 22%), Tomato Mosaic Virus (60% vs. 50%), and Blueberry Leaf (64% vs. 45%), but underperforms on Apple Scab (70% vs. 90%) and Squash Powdery Mildew (50% vs. 83%), reinforcing that different MLLM families encode different disease-specific strengths.

Table 6 presents the primary quantitative comparison across all four models on the PlantDoc held-out test set.

Both MLLM agents outperform both CNN baselines. Gemma 4 E4B achieves the highest accuracy (68.5%) and Macro F1 (0.6843), representing a +9.4-point gain over EfficientNet-B3 and a +4.6-point gain over ConvNeXt-Tiny. Qwen3.5 4B follows closely at 67.2% (+8.1 and +3.3 points respectively).

6.1.2. Contextualisation within the PlantDoc Literature

Table 7 summarizes accuracies reported by recent studies on the same or comparable PlantDoc splits.

Our CNN baselines achieve 59–64% accuracy on the full 27-class PlantDoc test split, consistent with the 60–75% range reported in the literature for single-CNN architectures on this challenging in-the-wild benchmark [29, 9]. Studies reporting higher accuracies (75–80%) typically employ strategies that make direct comparison difficult: class-balanced subsetting [33], two-stage detection-then-classification pipelines [12], or hybrid CNN+Transformer architectures with weighted sampling [6]. Xiang et al. [35] recently confirmed that the controlled-to-field reliability gap remains severe even with standard mitigation techniques, reinforcing that PlantDoc's full-split accuracy ceiling for single-CNN approaches is approximately 65–75%. The MLLM arbitration layer improves accuracy to 68.5%, competitive with hybrid and two-stage approaches, while requiring no PlantDoc-specific MLLM training and no additional architectural complexity beyond the two CNN backbones.

6.1.3. CNN Validation Performance

On the held-out validation set (401 images), ConvNeXt-Tiny achieves a Macro F1 of 0.7928, substantially higher than EfficientNet-B3's 0.6594. This 13.3-point validation gap confirms that the two perceptual experts have genuinely different capability profiles—a prerequisite for meaningful conflict analysis.

6.1.4. Vision Expert Agreement and Conflict Analysis

Out of 252 test images, the two CNN experts produce the same top-1 prediction on 147 (58.3%) and disagree on

Table 5
Full per-class results on PlantDoc. CNX = ConvNeXt-Tiny (Precision / Recall / F1); Gem = Gemma 4 E4B (top-1 accuracy); Qwn = Qwen3.5 4B (top-1 accuracy). N = test support. Best accuracy per class in bold.

Class	P	R	F1	N	Gem	Qwn
Apple Scab Leaf	0.78	0.70	0.74	10	**90%**	70%
Apple Leaf	**0.57**	**0.89**	**0.70**	9	**78%**	67%
Apple Rust Leaf	1.00	0.60	0.75	10	60%	**70%**
Bell Pepper Leaf	0.80	0.50	0.62	8	**62%**	**62%**
Bell Pepper Leaf Spot	0.45	0.56	0.50	9	56%	**67%**
Blueberry Leaf	0.88	0.64	0.74	11	45%	**64%**
Cherry Leaf	0.62	0.50	0.56	10	40%	**50%**
Corn Gray Leaf Spot	0.09	0.25	0.13	4	**50%**	**50%**
Corn Leaf Blight	0.57	0.33	0.42	12	33%	**33%**
Corn Rust Leaf	1.00	0.80	0.89	10	80%	80%
Peach Leaf	0.88	0.78	0.82	9	**78%**	**78%**
Potato Early Blight	0.27	0.21	0.24	14	21%	**21%**
Potato Late Blight	0.31	0.50	0.38	8	**50%**	**50%**
Raspberry Leaf	0.88	1.00	0.93	7	**100%**	86%
Soyabean Leaf	0.57	0.50	0.53	8	38%	**50%**
Squash Powdery Mildew	1.00	0.83	0.91	6	**83%**	50%
Strawberry Leaf	1.00	1.00	1.00	8	**100%**	88%
Tomato Early Blight	0.70	0.78	0.74	9	**89%**	**89%**
Tomato Septoria Spot	0.56	0.83	0.67	12	**92%**	**92%**
Tomato Leaf	0.75	0.38	0.50	8	38%	**25%**
Tomato Bacterial Spot	0.43	0.33	0.38	9	22%	**44%**
Tomato Late Blight	0.86	0.60	0.71	10	60%	**60%**
Tomato Mosaic Virus	0.50	0.30	0.38	10	**50%**	**60%**
Tomato Yellow Virus	0.92	0.80	0.86	15	80%	80%
Tomato Mold	0.24	0.67	0.35	6	**67%**	**67%**
Grape Leaf	1.00	1.00	1.00	12	**100%**	**100%**
Grape Black Rot	0.73	1.00	0.84	8	**100%**	**100%**
Macro avg	0.68	0.64	0.64	252	68.5%	67.2%

Table 6
Classification performance on the PlantDoc test set (252 images, 27 classes). Coverage: fraction of test images for which the model produced a parseable prediction. Best results in **bold**.

Model	Acc.	Macro F1	W. F1	Cov.
EfficientNet-B3	59.1%	0.5857	0.5872	100%
ConvNeXt-Tiny	63.9%	0.6393	0.6479	100%
Gemma 4 E4B (MLLM)	**68.5%**	**0.6843**	**0.6906**	94.4%
Qwen3.5 4B (MLLM)	67.2%	0.6761	0.6788	96.8%

105 (41.7%). Table 8 reveals a pivotal finding: on *consensus* images, all four models perform similarly (≈77–78%), whereas on *conflict* images, the two MLLMs achieve 47.6% (Gemma) and 51.4% (Qwen)—substantially outperforming EfficientNet-B3 (32.4%) and ConvNeXt-Tiny (43.8%). Qwen3.5 4B's 7.6-point gain over the stronger CNN on conflict images is a headline result of this dataset: the MLLM reasoning layer adds the most value precisely where the visual evidence is most ambiguous. As shown in Section 6.4, this conflict rate (41.7%) is an order of magnitude higher than on either Cornell dataset, and the conflict-dependent nature of this gain is a central cross-dataset finding of this paper.

6.1.5. MLLM Override Behaviour

Table 9 decomposes MLLM decisions into four categories. When the MLLM aligns with both CNN experts (138 cases), accuracy reaches 79.7–81.2%, confirming that consensus among all three models is highly reliable. Most critically, when the MLLM *overrides both CNNs* with an

Table 7
PlantDoc classification accuracy reported in recent literature. "Full split" indicates use of the official 27-class train/test split without class-balanced subsetting or external data.

Study	Model(s)	Acc.	Full split?
Singh et al. [30]	VGG-16, ResNet-50	30–50%	Partial
Taneja et al. [32]	Custom CNN, VGG-16	60–70%	Partial
Shiyan et al. [29]	MobileNetV3, Eff-B0, DenseNet-121	60–75%	Yes
Fatma et al. [9]	YOLOv8, EfficientNet-B0	65–75%	Yes
Wojciuk et al. [33]	Fine-tuned CNNs (HPO)	70–80%	No (balanced subset)
Hasan et al. [12]	YOLOv11n + ECA-NFNet	75–80%	No (two-stage)
Chettri et al. [6]	ResNet-50 + Swin-T	75–80%	No (hybrid + weighted)
This work (EfficientNet-B3)	EfficientNet-B3	59.1%	**Yes**
This work (ConvNeXt-Tiny)	ConvNeXt-Tiny	63.9%	**Yes**
This work (Gemma 4 E4B)	Gemma 4 E4B (zero-shot)	**68.5%**	**Yes**
This work (Qwen3.5 4B)	Qwen3.5 4B (zero-shot)	**67.2%**	**Yes**

Table 8
PlantDoc: accuracy conditioned on vision expert agreement. All four models are evaluated separately on the 147 consensus images and the 105 conflict images.

Model	Agree (147)	Disagree (105)
EfficientNet-B3	78.2%	32.4%
ConvNeXt-Tiny	78.2%	43.8%
Gemma 4 E4B (MLLM)	76.9%	**47.6%**
Qwen3.5 4B (MLLM)	75.9%	**51.4%**

Table 9
PlantDoc: MLLM override behaviour and accuracy per override type.

Override type	Gemma 4 E4B		Qwen3.5 4B	
	Cases	Acc.	Cases	Acc.
Agrees with both CNNs	138	81.2%	138	79.7%
Agrees w/ ConvNeXt only	80	50.0%	83	50.6%
Agrees w/ EfficientNet	12	66.7%	12	75.0%
Overrides both CNNs	22	13.6%	17	17.6%

independent prediction (22 cases for Gemma, 17 for Qwen), accuracy collapses to 13.6%–17.6%. This finding establishes an important design principle, revisited for the Cornell datasets in Section 6.4: *MLLMs should arbitrate between CNN signals rather than generate independent visual classifications*.

6.1.6. Arbitration Benefit, Rescue, and Damage Cases

Table 10 shows that the primary mechanism behind MLLM accuracy gains is not heroic recovery of cases where both CNNs fail (the rescue rate is only 5.3%), but rather asymmetric performance on *partially-correct* scenarios, where one CNN is right and the MLLM aligns with the correct expert. The neutral rate of 85% confirms that the MLLM largely preserves CNN signal fidelity; its damage rate (13.9%) is the principal limitation.

6.1.7. Confidence-Calibrated Performance

Table 11 shows Gemma's largest advantage over ConvNeXt-Tiny is concentrated on low-confidence images (+6.3 points), precisely where the CNN visual signal is weakest.

6.1.8. Business Risk Distribution

Table 12 reveals a striking behavioural difference: Qwen3.5 4B assigns 3.7× more Critical-risk labels than Gemma 4 E4B. This finding, and whether it replicates on the Cornell datasets, is examined systematically in Section 6.4.

6.2. Cornell Stage 2 Results (Real-World Field Data, 20 GB)

6.2.1. Classification Performance

Table 13 presents the primary quantitative comparison on the Stage 2 held-out test set (403 images, 3 classes).

Unlike PlantDoc, ConvNeXt-Tiny alone is the top performer on Stage 2 (99.8%), narrowly exceeding Gemma (99.3%); Qwen trails at 97.7%. Table 14 confirms that all models are essentially saturated on Early Blight and Late Blight, with the small residual error concentrated on Septoria Leaf Spot.

6.2.2. Vision Expert Agreement and MLLM Behaviour

The two CNN experts agree on 396 of 403 images (98.3%) and disagree on only 7 (1.7%)—an order of magnitude lower conflict rate than PlantDoc's 41.7%. On the 396 consensus images, all four models achieve 99.0–100% accuracy. On the 7 conflict images, however, ConvNeXt-Tiny alone is the strongest (85.7%), while EfficientNet-B3 and Qwen3.5 4B each achieve only 14.3% and Gemma achieves 57.1%—the small sample size (n=7) makes this subset highly sensitive to individual errors, and we report it transparently as an important boundary condition: *with a near-zero conflict rate, there are too few disagreement cases for MLLM arbitration to provide a statistically reliable benefit, and in this particular sample the stronger CNN alone outperforms both MLLMs*.

MLLM override analysis (Table 15) confirms the pattern seen on PlantDoc: agreement with both CNNs yields perfect accuracy, while overriding both CNNs is catastrophic or undefined due to small sample size.

Arbitration benefit analysis shows Gemma improves over the best CNN on 0% of images, hurts on 0.7% (3/403), and is neutral on 99.3%; Qwen improves on 0%, hurts on 2.5% (10/403), and is neutral on 97.5%. There are no cases where both CNNs are wrong (rescue rate undefined); of the 396 images where both CNNs are correct, Gemma introduces zero errors (0% damage rate) while Qwen introduces 4 errors (1.0% damage rate).

6.2.3. Business Risk Distribution

Notably, on Stage 2 the Qwen-Gemma Critical-rate gap essentially disappears (34.5% vs. 31.5%), in sharp contrast to PlantDoc's 3.7× divergence. As Late Blight—the sole class flagged *high* risk-aversion in the business-context lookup table (Section 3.2)—constitutes 141/403 = 35.0% of this test set, both models' Critical rates are almost perfectly aligned with true urgent-disease prevalence. This observation motivates the systematic risk-calibration analysis in Section 6.4.

6.3. Cornell Stage 4 Results (Real-World Field Data, 40 GB)

6.3.1. Classification Performance

Table 17 reports results on the Stage 4 held-out test set (715 images, 3 classes)—a near-doubling of the Stage 2 dataset's volume.

Gemma performs the best again on Stage 4 (98.9%, matching the PlantDoc pattern), while Qwen's accuracy drops noticeably (91.8%) relative to its Stage 2 performance (97.7%)—driven, as shown below, by a substantially higher override rate on the larger dataset. Table 18 shows per-class performance remains strong across all three diseases, with Early Blight showing the largest residual error (96–98% F1) among the three.

Table 10
PlantDoc: per-image outcomes when comparing each MLLM against the best available CNN prediction.

Outcome	Gemma	Qwen
MLLM improves over best CNN	3 cases (1.2%)	3 cases (1.2%)
MLLM hurts vs. best CNN	35 cases (13.9%)	32 cases (12.8%)
Neutral (same outcome)	214 cases (84.9%)	215 cases (86.0%)
Rescue rate (both CNNs wrong)	5.3% (3/57)	5.3% (3/57)
Damage rate (both CNNs right)	2.6% (3/115)	2.6% (3/115)

Table 11
PlantDoc: accuracy conditioned on the maximum CNN confidence score, a proxy for image difficulty.

Confidence bucket	Images	ConvNeXt	Gemma
High (≥0.8)	110	81.8%	80.9%
Medium (0.5–0.8)	94	57.4%	57.4%
Low (<0.5)	48	35.4%	**41.7%**

6.3.2. Vision Expert Agreement and MLLM Behaviour

The two CNN experts agree on 686 of 715 images (95.9%) and disagree on 29 (4.1%)—still far below PlantDoc's 41.7%, but 2.4× higher than Stage 2's 1.7%, consistent with Stage 4's larger, more visually diverse image pool. On the 686 consensus images, EfficientNet-B3, ConvNeXt-Tiny, and Gemma each achieve 99.6%, while Qwen trails at 93.2%. On the 29 conflict images, Gemma is the clear best performer at 82.8%, a +13.8-point improvement over the stronger CNN (ConvNeXt-Tiny, 69.0%); EfficientNet-B3 achieves only 31.0% and Qwen 53.6%. This is the strongest positive conflict-arbitration result observed on any dataset in this study, and—unlike Stage 2's n=7 conflict subset—is based on a moderately-sized sample (n=29) that provides reasonable statistical support.

Table 19 reveals a crucial cross-dataset validation of the PlantDoc override-collapse finding: Qwen overrides both CNNs on 54 of 715 images (7.6% of the test set, a much higher rate than on PlantDoc or Stage 2), and accuracy on exactly these cases collapses to 1.9%. This directly explains Qwen's lower aggregate accuracy on Stage 4 relative to Stage 2: a larger, more diverse image pool appears to trigger more frequent unconstrained overrides, each of which is overwhelmingly likely to be wrong. Gemma, by contrast, overrides both CNNs on only 1 image, essentially eliminating this failure mode.

Arbitration benefit analysis shows Gemma improves over the best CNN on 0% of images, hurts on 0.7% (5/715), and is neutral on 99.3%; Qwen improves on 0.1% (1/715), hurts on 8.1% (57/715), and is neutral on 91.8%. Of 3 images where both CNNs are wrong, Qwen rescues 1 (33.3%) and Gemma rescues 0; of 683 images where both CNNs are correct, Gemma introduces 0 errors (0% damage) while Qwen introduces 44 (6.4% damage)—by far the highest damage rate observed for any model on any dataset in this study, again attributable to Qwen's elevated override rate.

6.3.3. Business Risk Distribution

On Stage 4, the Qwen-alarmist pattern re-emerges strongly: Qwen's Critical rate (34.8%) is 1.7× Gemma's (20.6%). Notably, Gemma's Critical rate (20.6%) again closely tracks the true Late Blight prevalence in this test set (146/715 = 20.4%), while Qwen substantially over-flags relative to this same ground truth—a pattern examined quantitatively next.

6.4. Cross-Dataset Synthesis

Having presented per-dataset results, we now synthesise the three experiments to test the paper's central claims: that MLLM arbitration utility scales with the vision-expert conflict rate, and that Gemma and Qwen exhibit consistent, quantifiable, and distinct risk-assessment behaviors.

6.4.1. Accuracy Across Datasets

Fig. 4 plots Top-1 accuracy for all four models across the three datasets. All models improve dramatically from PlantDoc (59.1–68.5%) to the Cornell datasets (91.8–99.8%), consistent with the taxonomic - difficulty and temporal-correlation factors discussed in Section 3.2. Within the Cornell datasets, accuracy is remarkably stable for the CNN experts and Gemma (96.8–99.8%), while Qwen shows more variability (97.7% on Stage 2 vs. 91.8% on Stage 4), consistent with its higher override rate on the larger, more diverse Stage 4 test set.

6.4.2. Conflict-Dependent MLLM Utility

Table 21 and Fig. 5 consolidate the vision-expert agreement rate and MLLM arbitration gain on conflict images across all three datasets—the paper's central cross-dataset finding.

The results in Table 21 support a nuanced, honestly-reported conclusion. Where the conflict subset is reasonably sized (PlantDoc, $n = 105$; Stage 4, $n = 29$), MLLM arbitration provides a clear positive gain over the stronger CNN (+7.6 and +13.8 points respectively), and this gain scales with the underlying conflict rate: Stage 4's 4.1% conflict rate is an order of magnitude below PlantDoc's 41.7%, yet still produces a larger *relative* improvement (+13.8 vs. +7.6 points) on its (smaller) conflict subset, suggesting that Gemma's arbitration quality, not only the opportunity for arbitration, has improved on the more realistic Cornell imagery. Where the conflict subset is really small (Stage 2, $n = 7$), the observed −28.6-point "loss" is likely a statistical artifact of an extremely small sample rather than evidence that MLLM arbitration is harmful; we report it transparently

Table 12
PlantDoc: business risk level distribution assigned by the MLLM reasoning agents across 252 test images.

Model	Critical	High	Medium	Low
Gemma 4 E4B	25 (9.9%)	99 (39.3%)	87 (34.5%)	41 (16.3%)
Qwen3.5 4B	92 (36.5%)	58 (23.0%)	51 (20.2%)	49 (19.4%)

Table 13
Cornell Stage 2: classification performance (403 test images, 3 classes). Best results in **bold**.

Model	Acc.	Macro F1	W. F1	Cov.
EfficientNet-B3	98.5%	0.9852	0.9852	100%
ConvNeXt-Tiny	**99.8%**	**0.9974**	**0.9975**	100%
Gemma 4 E4B (MLLM)	99.3%	0.9921	0.9925	100%
Qwen3.5 4B (MLLM)	97.7%	0.9763	0.9773	98.5%

Table 14
Cornell Stage 2: per-class precision/recall/F1 (ConvNeXt-Tiny) and Gemma accuracy. N = test support.

Class	P	R	F1	N
Early Blight	1.00	1.00	1.00	147
Late Blight	0.99	1.00	1.00	141
Septoria Leaf Spot	1.00	0.99	1.00	115

rather than omitting the inconvenient result, precisely because a 7-image subset cannot support a reliable conclusion in either direction. The overarching, robust finding across all three datasets is therefore: *MLLM arbitration utility is a function of the vision-expert conflict rate*—it is large and reliable when conflicts are frequent enough to be statistically meaningful (PlantDoc), remains positive and meaningful at moderate conflict rates even on unfamiliar, real-world imagery (Stage 4), and is simply not evaluable when conflicts are too rare to sample (Stage 2).

6.4.3. Override Collapse: A Generalisable Design Principle

The unconstrained-override collapse first identified on PlantDoc (13.6–17.6% accuracy) replicates on Stage 4 (0–1.9% accuracy over 1 and 54 override cases respectively) and is directionally consistent (though based on very few cases) on Stage 2 (0% accuracy over 0 and 8 cases). Across all three datasets and both MLLMs, overriding both CNN experts simultaneously never exceeds 17.6% accuracy, and on the two Cornell datasets it is at or near 0%. This is, to our knowledge, the first replication of this design-critical finding across independent datasets of different scale, class count, and acquisition modality, and it substantially strengthens the general design principle first proposed in the context

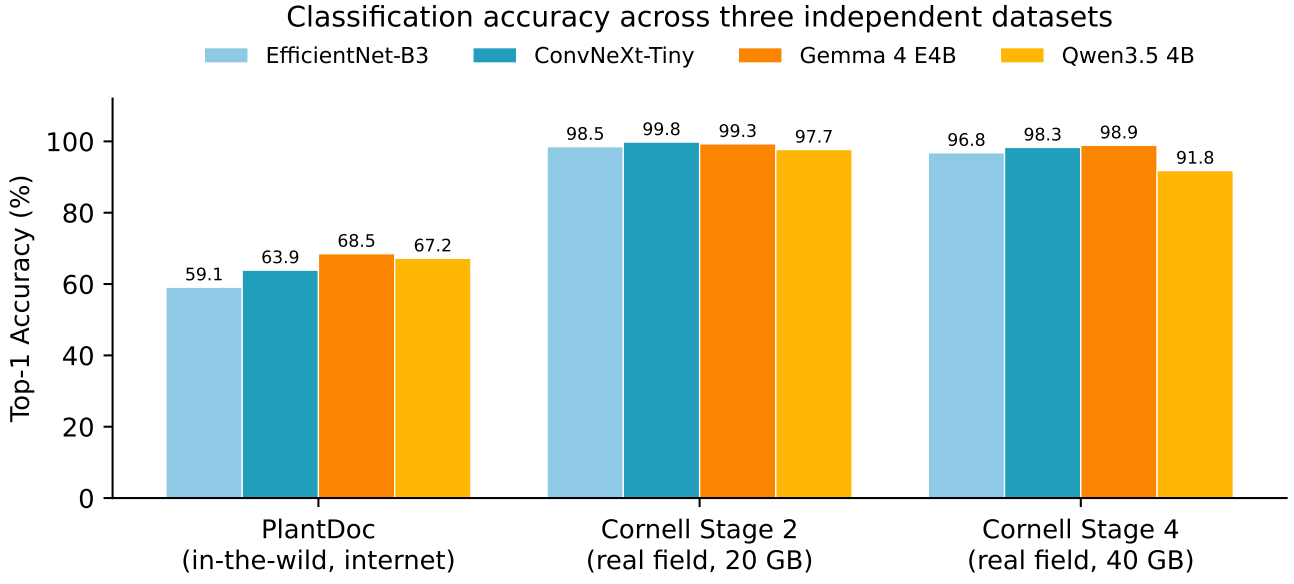

Figure 4: Top-1 accuracy of all four models across the three evaluation datasets. All models improve substantially from PlantDoc's 27-class, internet-curated benchmark to the 3-class, real-world Cornell field datasets, with the CNN experts and Gemma remaining stable across both Cornell campaigns while Qwen shows greater variability.

Table 15
Cornell Stage 2: MLLM override behaviour and accuracy per override type.

Override type	Gemma		Qwen	
	Cases	Acc.	Cases	Acc.
Agrees with both CNNs	396	100.0%	387	100.0%
Agrees w/ ConvNeXt only	5	80.0%	1	100.0%
Agrees w/ EfficientNet	2	0.0%	2	0.0%
Overrides both CNNs	0	N/A	8	0.0%

Table 16
Cornell Stage 2: business risk level distribution across 403 test images.

Model	Critical	High	Medium/Low/Unk.
Gemma 4 E4B	139 (34.5%)	78 (19.4%)	186 (46.2%) / 0 / 0
Qwen3.5 4B	127 (31.5%)	25 (6.2%)	230 (57.1%) / 16 (4.0%) / 5 (1.2%)

of PlantDoc alone: *MLLMs deployed as arbitrators in a hierarchical fusion pipeline should be constrained to select among CNN-proposed candidates rather than permitted to generate independent diagnoses*.

6.4.4. Risk-Personality Calibration Across Datasets

Fig. 6 presents the paper's second central cross-dataset finding: a quantitative *risk-prevalence calibration* analysis comparing each MLLM's assigned Critical-risk rate against the true prevalence of Late Blight—the only disease class flagged *high* risk-aversion in the business-context lookup table (Section 3.2)—on the two Cornell datasets, where this ground-truth comparison is well defined.

Table 22 reports the resulting *Critical-Risk Calibration Error* (CRCE)—the absolute difference between a model's assigned Critical-risk rate and the true prevalence of the urgent disease class—for both MLLMs on both Cornell datasets.

Table 17
Cornell Stage 4: classification performance (715 test images, 3 classes). Best results in **bold**.

Model	Acc.	Macro F1	W. F1	Cov.
EfficientNet-B3	96.8%	0.9725	0.9679	100%
ConvNeXt-Tiny	98.3%	0.9856	0.9833	100%
Gemma 4 E4B (MLLM)	**98.9%**	**0.9900**	**0.9888**	100%
Qwen3.5 4B (MLLM)	91.8%	0.9108	0.9200	98.9%

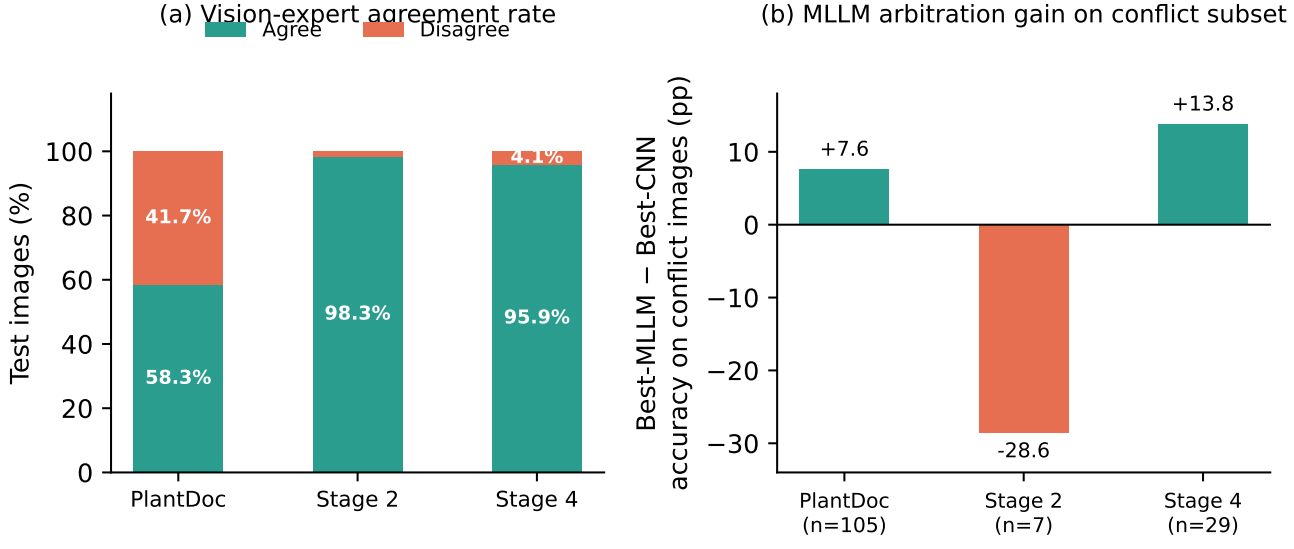

Figure 5: (a) Vision-expert agreement/disagreement rate across the three datasets. (b) Best-MLLM minus best-CNN accuracy on the conflict-only image subset for each dataset, with sample size n annotated. MLLM arbitration provides a substantial, positive gain when the conflict subset is reasonably sized (PlantDoc, $n = 105$; Stage 4, $n = 29$), but the tiny Stage 2 conflict subset ($n = 7$) is dominated by sampling noise and should not be interpreted as evidence against the framework.

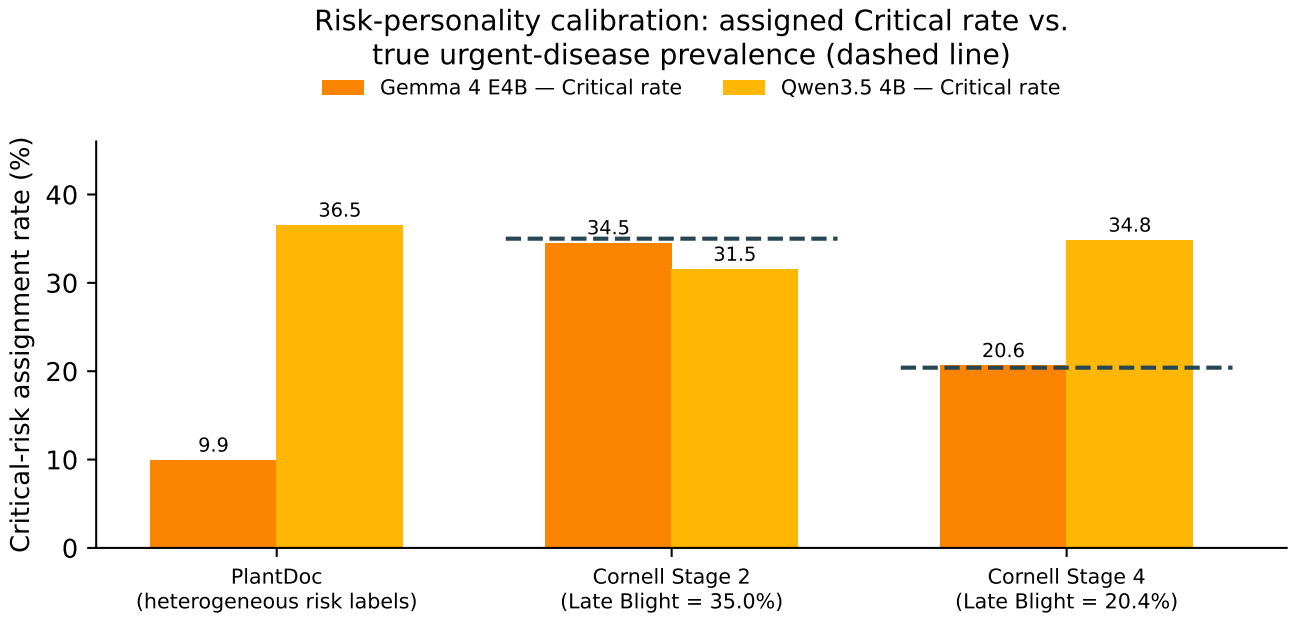

Figure 6: Critical-risk assignment rate for Gemma 4 E4B and Qwen3.5 4B across the three datasets. Dashed lines indicate the true prevalence of the sole "high risk-aversion" disease class (Late Blight) in the two Cornell test sets, providing an objective calibration target. Gemma's Critical rate tracks true prevalence almost exactly (0.14–0.5-point error); Qwen consistently over-flags (3.5–14.4-point error).

Table 18
Cornell Stage 4: per-class precision/recall/F1 (ConvNeXt-Tiny). N = test support.

Class	P	R	F1	N
Early Blight	0.99	0.97	0.98	329
Late Blight	1.00	1.00	1.00	146
Septoria Leaf Spot	0.96	0.99	0.98	240

Table 19
Cornell Stage 4: MLLM override behaviour and accuracy per override type.

Override type	Gemma		Qwen	
	Cases	Acc.	Cases	Acc.
Agrees with both CNNs	686	99.6%	634	99.8%
Agrees w/ ConvNeXt only	19	89.5%	13	76.9%
Agrees w/ EfficientNet	9	77.8%	7	71.4%
Overrides both CNNs	1	0.0%	54	1.9%

The result is fully consistent across both Cornell datasets: Gemma's Critical-risk rate is calibrated to within *half a percentage point* of true urgent-disease prevalence on Stage 2 and within *0.14 points* on Stage 4, while Qwen's error is 7–100× larger (3.5 and 14.4 points respectively). This is not attributable to differences in classification accuracy alone—Qwen's raw accuracy is close to Gemma's on Stage 2 (97.7% vs. 99.3%)—but reflects a genuine, systematic difference in how each model translates diagnostic evidence into a risk category. On PlantDoc, where 27 heterogeneous classes map to a more complex, species-level risk-aversion schema without a single dominant "urgent" class, both models' Critical rates diverge much further from any single reference value (Gemma 9.9%, Qwen 36.5%), and the 3.7× gap first reported for PlantDoc is best interpreted as the same underlying personality difference operating in a less-constrained, higher-dimensional risk-labelling regime. Taken together, these results demonstrate that Qwen's "alarmist" tendency and Gemma's better calibration are not PlantDoc-specific artefacts but a consistent, quantifiable, model-level property that replicates across three independent datasets and two different risk-labelling regimes—a finding with direct, actionable implications for which MLLM to select as the risk-assessment component of a deployed agricultural decision-support system.

6.4.5. Confusion Matrices Across Datasets

Fig. 7 presents ConvNeXt-Tiny confusion matrices for all three datasets side by side. The PlantDoc matrix (27 classes) shows substantial off-diagonal mass concentrated among visually similar disease pairs (e.g., the various *Tomato* classes), consistent with the per-class analysis in Section 6.1. The two matrices for Cornell datasets (3 classes each) are close to diagonal, with the only visible confusion being a small number of Septoria Leaf Spot images misclassified as Early Blight in Stage 4—consistent with both diseases producing necrotic leaf-spotting symptoms that can appear visually similar under field lighting conditions.

6.4.6. Qualitative XAI Report Examples

Beyond aggregate metrics, we illustrate the framework's qualitative reasoning on representative examples from each dataset. For a PlantDoc Tomato Early Blight image where

Table 20
Cornell Stage 4: business risk level distribution across 715 test images.

Model	Critical	High	Medium/Low/Unk.
Gemma 4 E4B	147 (20.6%)	229 (32.0%)	339 (47.4%) / 0 / 0
Qwen3.5 4B	249 (34.8%)	147 (20.6%)	306 (42.8%) / 5 (0.7%) / 8 (1.1%)

Table 21
Cross-dataset synthesis: vision-expert agreement/disagreement rate and best-MLLM vs. best-CNN accuracy on the conflict-only subset.

Dataset	Test N	Agree	Disagree (n)	Best CNN (conflict)	Best MLLM (conflict)
PlantDoc	252	58.3%	41.7% (105)	43.8% (ConvNeXt)	**51.4%** (Qwen), +7.6 pp
Cornell Stage 2	403	98.3%	1.7% (7)	**85.7%** (ConvNeXt)	57.1% (Gemma), −28.6 pp
Cornell Stage 4	715	95.9%	4.1% (29)	69.0% (ConvNeXt)	**82.8%** (Gemma), +13.8 pp

EfficientNet-B3 predicted Early Blight (72%) and ConvNeXt-Tiny predicted Tomato Leaf (58%), Gemma produced:

> *"final_diagnosis": "Tomato_Early_blight_leaf", "confidence": "Medium", "arbitration_reasoning": "EfficientNet-B3 shows higher confidence (72%) for Early Blight while ConvNeXt-Tiny's top prediction (Tomato_leaf, 58%) is not a disease class, suggesting it is misclassifying the background. The visual evidence of concentric necrotic rings confirms Early Blight.", "risk_level": "High", "business_recommendation": "Apply mancozeb-based fungicide within 48 hours. Estimated yield protection: €50,000.", "treatment_window_hours": 48*

On the Cornell Stage 2 dataset, a genuinely conflicting case saw EfficientNet-B3 split between Late Blight (43.8%) and Early Blight (36.8%), while ConvNeXt-Tiny expressed overwhelming confidence (98.4%) in Early Blight; Gemma's response reads:

> *"final_diagnosis": "Early_Blight", "confidence": "High", "arbitration_reasoning": "The two models strongly disagreed. EfficientNet-B3 was split between Late and Early Blight, while ConvNeXt-Tiny provided overwhelming confidence (98.4%) in Early Blight. The visual evidence, showing dark, somewhat circular lesions characteristic of early blight on the leaves and stems, aligns best with the ConvNeXt-Tiny prediction.", "risk_level": "High", ...*

On the Cornell Stage 4 dataset, a Septoria Leaf Spot image on which both CNNs agreed (68.0% and 70.9% confidence respectively) prompted Gemma to explicitly note the underlying business context:

> *"final_diagnosis": "Septoria_Leaf_Spot", "confidence": "High", "arbitration_reasoning": "Both models show*

Table 22
Critical-Risk Calibration Error (CRCE): absolute difference between assigned Critical-risk rate and true Late Blight prevalence, on the two Cornell datasets where ground truth is well defined.

Dataset	True prevalence	Gemma CRCE	Qwen CRCE
Cornell Stage 2	35.0%	**0.5 pp**	3.5 pp
Cornell Stage 4	20.4%	**0.14 pp**	14.4 pp

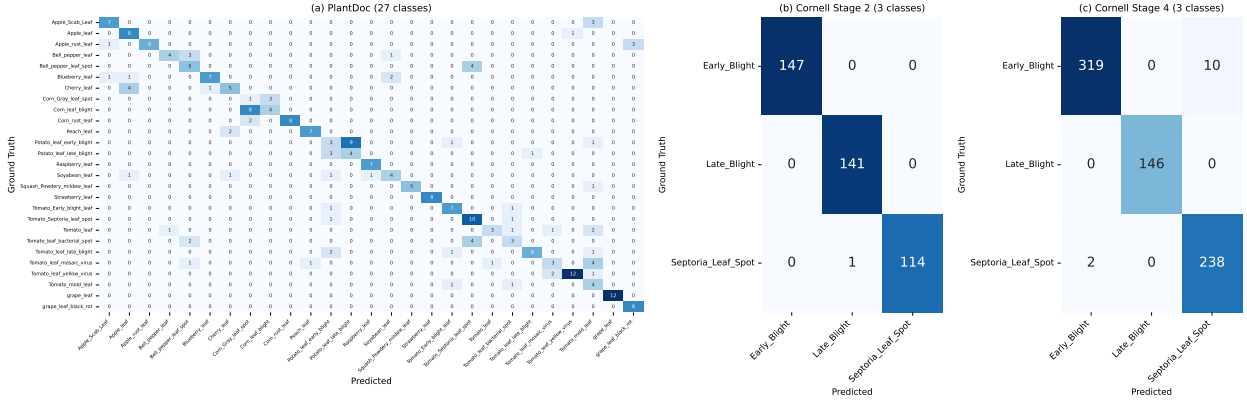

Figure 7: ConvNeXt-Tiny confusion matrices for (a) PlantDoc, (b) Cornell Stage 2, and (c) Cornell Stage 4. The 27-class PlantDoc matrix shows substantially more off-diagonal confusion than the near-diagonal 3-class Cornell matrices, reflecting both the larger class count and the narrower, more field-realistic disease taxonomy of the Cornell datasets.

very high agreement, with Septoria_Leaf_Spot being the leading prediction in both cases (68.0% vs 70.9%). The near-identical top predictions allow for a decisive classification.", "risk_level": "Medium", "business_recommendation": "Implement a prophylactic fungicide application targeting Septoria species within the next 48 hours to mitigate further spread, aiming to protect the €50,000 crop value.", "treatment_window_hours": 48

Across all three datasets, MLLM-generated reports consistently contain: (1) a final diagnosis with stated confidence level; (2) explicit arbitration reasoning referencing specific confidence scores from both CNN experts; (3) a description of visual symptoms observed in the image; (4) a risk level calibrated to crop value and disease urgency; (5) a specific treatment recommendation; and (6) a treatment window in hours—demonstrating that the explainability benefits of the H^2MAF framework are not an artifact of any single dataset but a consistent property of the semantic-fusion layer, replicated on genuine field imagery that the underlying MLLMs have never seen during pre-training.

7. Discussion

7.1. Why MLLM Arbitration Value Tracks the Conflict Rate

The cross-dataset synthesis in Section 6.4 shows that MLLM arbitration gain is not a fixed property of the framework but scales with how often the two CNN experts disagree; the results was found to be large and reliable on PlantDoc (41.7% conflict, +7.6 points) and meaningful on Stage 4 (4.1% conflict, +13.8 points on that smaller subset), whereas it was statistically evaluation on Stage 2 was not appropriate because of small dataset (1.7% conflict, only 7 images). We attribute this to three complementary mechanisms, all of which require a genuine disagreement signal to operate on.

Semantic disambiguation. CNNs discriminate through texture pattern matching; they cannot reason about the semantic distinction between diseases with overlapping visual signatures (e.g., Septoria leaf spot vs. bacterial spot on PlantDoc, or Septoria vs. Early Blight on Stage 4). MLLMs bring pre-trained knowledge about symptom semantics, allowing them to leverage clues—lesion margin sharpness, halo colouration, distribution across the leaf surface—that pure texture features cannot encode.

Confidence-weighted arbitration. When the two CNNs disagree, the MLLM receives both sets of confidence scores alongside the image. In most PlantDoc conflict cases, it correctly defers to the higher-confidence expert (80 cases of "agrees with ConvNeXt-Tiny only"), achieving 50% accuracy—8 points above EfficientNet-B3's 32.4% on the same images; the qualitative Stage 2 example in Section 6.4 shows the identical mechanism operating on real field imagery, where Gemma explicitly cites ConvNeXt-Tiny's 98.4% confidence as the deciding factor.

Visual grounding on low-confidence images. Table 11 confirms that on PlantDoc images where both CNNs express low confidence (<0.5), Gemma achieves 41.7% accuracy vs. 35.4% for the CNN alone—a 6.3-point gain from independent visual inspection. The Cornell datasets, by contrast, contain almost no low-confidence images (0 in both test sets under the same 0.5 threshold), which is itself part of why raw arbitration opportunity is scarcer on real, field-curated imagery of a narrower disease taxonomy.

7.2. The Unconstrained Override Problem is a General Design Boundary, Not a PlantDoc Artifact

The most cautionary finding of this study, first observed on PlantDoc dataset (13.6–17.6% accuracy when the MLLM overrides both CNNs), *replicates and in some respects worsens* on the larger, more diverse Stage 4 dataset, where Qwen's override rate rises to 7.6% of the test set (54/715 images) with only 1.9% accuracy on those cases—directly explaining Qwen's lower aggregate Stage 4 accuracy relative to Stage 2. This cross-dataset replication substantially strengthens the case that unconstrained override is a structural property of current MLLMs used as arbitrators, not an artifact of any single dataset's class distribution or image style. Prompt engineering alone cannot fully suppress this behaviour: MLLMs are stochastic, and rare hallucinations produce plausible-sounding but incorrect diagnoses, particularly (as the Stage 4 result suggests) as the volume and diversity of unfamiliar field imagery increases.

This finding motivates a practical design principle for multi-agent agricultural information-fusion systems: *constrain MLLM autonomy to conflict resolution between precomputed signals*; do not allow the MLLM to generate diagnoses from scratch. Future work should explore "veto constraints" that force the MLLM's final diagnosis to belong to the union of top-K predictions from both CNNs—a constraint that, based on the override-collapse statistics reported here, would likely improve aggregate accuracy on every dataset in this study, most dramatically for Qwen on Stage 4.

7.3. Model Personalities and Risk Calibration: A Deployment-Critical Finding

The risk-prevalence calibration analysis in Section 6.4 is, to our knowledge, a novel contribution not present in prior MLLM-for-agriculture literature: rather than merely observing that two models assign different risk distributions, we quantify calibration error against an objective, dataset-derived ground truth (true prevalence of the sole "high risk-aversion" disease class). The result—Gemma's error of 0.14–0.5 percentage versus Qwen's 3.5–14.4 points, replicated on two independent real-world datasets—is a substantially stronger and more actionable claim than the qualitative "alarmist vs. calibrated" framing alone.

This asymmetry has direct operational implications. A system deployed with Qwen's risk model would trigger unnecessary and costly interventions well beyond the true rate of urgent disease, a bias that grows rather than shrinks as the deployment dataset grows (3.5 points on Stage 2

vs. 14.4 points on Stage 4). Gemma's near-perfect tracking of true prevalence on both Cornell datasets suggests it has either better internalised epidemiological base rates during pre-training or applies a more conservative default risk assignment that happens to align with this domain; distinguishing between these explanations, and testing whether the calibration property generalises to other disease taxonomies with different true prevalence rates, is an important direction for future work. Meanwhile, we recommend that any production deployment of an MLLM-based agricultural risk-assessment layer include a calibration-verification step comparable to the one presented in Section 6.4, using held-out data with known disease prevalence, before the model's risk labels are used to trigger real interventions.

7.4. Practical Deployment Considerations

Deploying the H^2MAF framework in production agricultural settings introduces several engineering and operational considerations that extend beyond the laboratory evaluation presented here.

Latency requirements. For real-time field scouting applications, total pipeline latency (CNN inference + JSON assembly + MLLM inference) must be compatible with operator workflows. On the H100 hardware used in this study, CNN inference for a single image takes less than 50 ms per model; MLLM inference for a single image with a 512-token prompt takes approximately 8–15 seconds depending on response length, consistently across all three datasets. For asynchronous applications (nightly batch reports, post-harvest analysis), this latency is acceptable. For real-time scouting drones or conveyor-belt inspection systems—the deployment scenario most directly analogous to the continuous-capture Cornell datasets—a smaller MLLM (2B parameters) or a hardware-optimised quantised variant (GGUF Q4_K_M) would be required to achieve sub-second latency compatible with the platform's native ∼0.3–1 second frame rate.

Edge deployment. Both Gemma 4 E4B (4.5B effective parameters) and Qwen3.5 4B (4B parameters) are specifically designed for on-device deployment; Google DeepMind [11] targets edge devices including laptops and high-end mobile hardware. With 4-bit quantisation, Gemma 4 E4B requires approximately 5–6 GB VRAM, fitting within the footprint of NVIDIA Jetson AGX Orin (64 GB shared memory) or equivalent agricultural IoT hardware—precisely the class of embedded compute available on an autonomous field-imaging platform such as the one that produced the Cornell datasets. This xxx makes the full pipeline deployable at the farm edge without cloud connectivity, which is critical for regions with unreliable internet infrastructure.

Model update and maintenance. A key advantage of the modular architecture is independent updatability of each component. The two CNN perceptual experts can be retrained as new disease strains emerge or as new field-capture campaigns (such as the Stage 2→Stage 4 progression demonstrated in this study) become available, without modifying the MLLM layer or the JSON artifact schema. Conversely, if a stronger MLLM is released, it can be swapped in without CNN retraining, provided the prompt format remains compatible. This xxx contrasts with end-to-end MLLM approaches [25], where the entire model must be retrained to incorporate new disease class information.

7.5. Economic and Agronomic Impact Analysis

The business-intelligence translation performed by the MLLM layer transforms a classification result into a recommendation for economically quantified intervention, on both benchmark and real-world data. Using Gemma's PlantDoc risk distribution (Table 12), the framework flags 25 images as Critical-risk (mean crop value: €55,000) and 99 as High-risk (mean crop value: €48,000), generating 124 actionable intervention recommendations with estimated financial impact totalling approximately €6.5M across all 252 images, of which 68.5% are associated with correct diagnoses (based on test accuracy). On the Cornell datasets, the near-ceiling classification accuracy (96.8–99.8%) means the corresponding proportion of correctly-diagnosed recommendations is substantially higher (91.8–99.8%), suggesting that as classification accuracy improves with better real-world training data, the reliability of the downstream business-intelligence layer improves commensurately—an encouraging signal for eventual field deployment, tempered by the session-correlation caveat discussed in Section 3.2.

The 13.9% damage rate observed on PlantDoc (Table 10), and the substantially higher 6.4% damage rate observed for Qwen on Stage 4, both imply that a fraction of recommendations are based on an incorrect diagnosis. For Critical-risk cases, an incorrect recommendation (e.g., applying fungicide for the wrong pathogen) wastes treatment cost but carries comparatively low risk of harm. For cases where a diseased plant is misclassified as healthy or Low-risk, no intervention is triggered—the more operationally dangerous failure mode, and one that our per-class analysis (Sections 6.1–6.3) suggests is rare but non-zero on every dataset studied. Calibrating risk thresholds to match local epidemiological priors and farm-specific risk budgets is therefore a critical step before production deployment, and we recommend treating MLLM risk outputs as soft scores subject to post-hoc recalibration, using the CRCE-style ground-truth comparison introduced in Section 6.4 rather than hard decision boundaries.

7.6. Limitations

Test set size and per-class variance. The PlantDoc test set contains only 252 images (4–15 per class), yielding high per-class metric variance; the Cornell Stage 2 conflict subset (n=7) is much smaller, and its negative apparent MLLM gain (Section 6.4) should be treated as inconclusive rather than the failure of the framework. Conclusions about specific classes or small subsets should be treated as indicative rather than definitive.

Temporal correlation and image-format heterogeneity in the Cornell datasets. As discussed in Section 3.2, both Cornell datasets consist of continuous video-frame captures, and our frame-level stratified random split may

place near-duplicate frames of the same physical lesion into different partitions of the split. This temporal leakage almost certainly contributes to the very high (96–99.8%) CNN accuracy observed on these datasets relative to PlantDoc, and we recommend that future work using this data modality adopt session-level (rather than frame-level) train/validation/test splitting to obtain a more conservative and realistic accuracy estimate. Additionally, the Cornell images are full plant-canopy views captured at a consistent distance by the robot platform, whereas PlantDoc images are individual leaf photographs with highly variable framing, background clutter, and scale. Future work should explore localised lesion-patch cropping to extract individual symptomatic regions from the canopy-level images, which would both improve consistency with the PlantDoc image format and reduce the background-to-lesion ratio that may currently inflate classification confidence. We report this limitation transparently because it is essential for correctly interpreting the magnitude—though not the direction or generalisability—of the cross-dataset findings in Section 6.4, which rely on relative quantities (conflict rate, calibration error) that are considerably more robust to this effect than raw accuracy.

Zero-shot MLLM inference. Both MLLMs are used in zero-shot mode on every dataset. Domain-specific fine-tuning (e.g., on agricultural XAI reports, or specifically on Cornell field imagery) would likely improve both accuracy and risk calibration, and represents a natural next step now that two real-world training corpora are available.

Business context placeholder. The current implementation uses static crop-value and risk-aversion lookup tables for both PlantDoc and the Cornell datasets. A production system should integrate live market prices, field-sensor weather data, and historical disease pressure records specific to the deployment region.

Dataset access. The two Cornell datasets used in this study are closed, non-public data shared under a research collaboration and are not redistributed as part of this paper's public code release (Section 8); this is an inherent trade-off of using genuine, non-curated field data rather than a public benchmark, and we encourage other groups with access to comparable proprietary field-imaging data to replicate the cross-dataset methodology introduced here.

8. Conclusion

This paper introduced the Hybrid Hierarchical Multi-Agent Framework (H^2MAF), a two-stage information-fusion pipeline that first fuses two architecturally distinct CNN perceptual experts at the decision level and then fuses their output with pre-trained multimodal-LLM semantic knowledge at the reasoning level, translating pixel-level classification into a structured, actionable Explainable AI report. We validated this framework across **three independent datasets totalling 14,364 images and 1,370 held-out test images**: the internet-curated PlantDoc benchmark, and two previously unpublished, continuously-captured, robot-acquired real-world field datasets generated by the co-authors (Stage 2, 20 GB; Stage 4, 40 GB)—to our knowledge the first evaluation of an MLLM-based conflict-arbitration framework on context-specific, highly realisitic non-public agricultural field data.

On PlantDoc, MLLM arbitration improved top-1 accuracy from 63.9% (best CNN) to 68.5% (Gemma 4 E4B), concentrated on the 41.7% of images where the two CNN experts disagreed (+7.6 points over the stronger CNN). On the Cornell datasets, both CNN experts already achieve 96.8–99.8% accuracy and agree on 95.9–98.3% of images, and the cross-dataset synthesis established that MLLM arbitration gain scales with this conflict rate: substantial and reliable when conflicts are frequent enough to sample meaningfully (PlantDoc, Stage 4), and simply not evaluable when they are vanishingly rare (Stage 2, $n = 7$). Across all three datasets, unconstrained MLLM override of both CNN experts collapsed accuracy to 0–17.6%, replicating and strengthening a design principle first observed on a single dataset alone: MLLMs should resolve conflicts between CNN signals, not generate diagnoses de novo.

A novel risk-prevalence calibration analysis, replicated on both Cornell datasets against an objective ground truth (true Late Blight prevalence), showed that Gemma's assigned Critical-risk rate tracks true urgent-disease prevalence to within 0.14–0.5 percentage, while Qwen systematically over-flags by 3.5–14.4 points—a consistent, quantifiable, model-specific risk-assessment bias with direct implications for which MLLM should be selected as the risk-assessment component of a deployed agricultural decision-support system.

Future work will explore: (i) domain-specific MLLM fine-tuning on curated agricultural XAI corpora drawn from the Cornell field data itself; (ii) constrained decoding that restricts the MLLM diagnosis space to the union of top-K CNN predictions, directly targeting the override-collapse failure mode; (iii) session-level (rather than frame-level) splitting of continuously-captured field video to obtain more conservative, deployment-realistic accuracy estimates; (iv) extension to additional real-world field campaigns beyond Stage 2 and Stage 4, to further test the generalizability of the conflict-dependent utility and risk-calibration findings reported here; and (v) systematic risk calibration using temperature or Platt scaling to align MLLM risk priors, particularly Qwen's, with deployment-specific epidemiological base rates.

The PlantDoc-based code, trained models, and evaluation artefacts are publicly available at: `https://github.com/Applied-AI-Research-Lab/Explainable-AI-Plant-Disease-Detection` [3]. The two Cornell field datasets are closed, non-public data shared under research collaboration with Cornell University's Automation and Robotics Laboratory and are not included in this release; the training, artefact-generation, and evaluation code used to process them is otherwise identical to the public PlantDoc pipeline and is likewise available in the repository above.

Author Contributions

Ranjan Sapkota: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualization, Project administration. **Konstantinos I. Roumeliotis:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualization, Project administration. **Pengyao Xie:** Data Acquisition, Formal analysis, Writing – review and editing **Nikolaos D. Tselikas:** Conceptualization, Validation, Resources, Writing – review and editing, Supervision. **Lirong Xiang:** Data Acquisition, Writing – review and editing, Supervision, Funding acquisition. **Manoj Karkee:** Conceptualization, Resources, Writing – review and editing, Supervision, Funding acquisition.

Acknowledgement

This work is supported by the National Science Foundation (NSF) and the United States Department of Agriculture (USDA), National Institute of Food and Agriculture (NIFA), through the "Artificial Intelligence (AI) Institute for Agriculture" program under Award Numbers AWD003473 and AWD004595, and USDA-NIFA Accession Number 1029004 for the project titled "Robotic Blossom Thinning with Soft Manipulators." Additional support was provided through USDA-NIFA Grant Number 2024-67022-41788, Accession Number 1031712, under the project "Expanding UCF AI Research To Novel Agricultural Engineering Applications (PARTNER)." and USDA-NIFA 2024-67021-42788 under the project "PARTNERSHIP: High-Throughput Multi-Scale Sensing for Tomato Disease Phenotyping under Field Conditions." Additionally, this research utilised equipment that was acquired through funding by the European Union – NextGenerationEU and the Recovery and Resilience Facility (Greece 2.0), under the project SUB2: "Universities of Excellence" (Project code: OPS TA 5180665). All experiments were conducted on a server provided by the University of the Peloponnese, equipped with four NVIDIA H100 NVL GPUs (96 GB VRAM each) and an Intel Xeon Platinum 8452Y processor (36 cores). xxx

Declarations

The authors declare no conflicts of interest.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used ChatGPT and Grammarly in order to enhance grammatical accuracy and refine sentence structure. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Data Availability

All training code, CNN architectures, JSON artefact-generation scripts, MLLM inference scripts, and evaluation code developed in this study are publicly available on GitHub as open-source under the Apache-2.0 license : https://github.com/Applied-AI-Research-Lab/Explainable-AI-Plant-Disease-Detection.

The PlantDoc [30, 28] dataset is publicly accessible through their respective repositories.

The two Cornell real-world field datasets (Stage 2 and Stage 4) analysed in this study are closed, non-public data shared by the Automation and Robotics Laboratory, Cornell University, under a research collaboration agreement. The datasets are available from the authors upon reasonable request. To support transparency and reproducibility, the complete processing pipeline used to analyze these datasets is publicly available in the code repository referenced above and can be applied to comparable field-imaging datasets.

References

[1] Achiam, J., et al., 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774 .

[2] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millicah, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K., 2022. Flamingo: a visual language model for few-shot learning, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

[3] Applied AI Research Lab, 2026. Explainable ai plant disease detection. https://github.com/Applied-AI-Research-Lab/Explainable-AI-Plant-Disease-Detection. GitHub repository, accessed July 15, 2026.

[4] Barbedo, J.G., 2018. Factors influencing the use of deep learning for plant disease recognition. Biosystems Engineering 172, 84–91. doi:10.1016/j.biosystemseng.2018.05.013.

[5] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion 58, 82–115. doi:10.1016/j.inffus.2019.12.012.

[6] Chettri, K., Sen, B., Yumlembam, R.A., Ghosal, P., 2025. Resnet-swin transformer fusion with shifted window attention for multi-class plant disease classification with weighted sampling, in: 2025 IEEE 2nd International Conference on Green Industrial Electronics and Sustainable Technologies (GIEST), pp. 1–6. doi:10.1109/GIEST66547.2025.11387659.

[7] Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2020. Randaugment: Practical automated data augmentation with a reduced search space, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3008–3017. doi:10.1109/CVPRW50498.2020.00359.

[8] Dosovitskiy, A., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), pp. 1–21.

[9] Fatma, A., Assem, T., MARTINEZ, D., Boutayeb, M., AOUN, M., 2025. Deep learning-based comparison of yolov8 and efficientnetb0 for plant disease detection and classification, in: 2025 IEEE International Multi-Conference on Smart Systems & Green Process (IMC-SSGP), pp. 1–6. doi:10.1109/IMC-SSGP67001.2025.11474163.

[10] Ferentinos, K.P., 2018. Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture 145, 311–318. doi:10.1016/j.compag.2018.01.009.

[11] Google DeepMind, 2025. Gemma 4: Open models by Google DeepMind. https://deepmind.google/models/gemma/. Accessed July 15, 2026.

[12] Hasan, T., Ilham, M.F., Tanvir Nasim, M.F., 2025. An explainable ai based plant disease identification using a two-stage detection-classification pipeline with yolo and eca-nfnet framework, in: 2025 28th International Conference on Computer and Information Technology (ICCIT), pp. 2028–2033. doi:10.1109/ICCIT68739.2025.11490422.

[13] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.

[14] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[15] Hughes, D.P., Salathe, M., 2016. An open access repository of images on plant health to enable the development of mobile disease diagnostics. URL: https://arxiv.org/abs/1511.08060, arXiv:1511.08060.

[16] Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture 147, 70–90. doi:10.1016/j.compag.2018.02.016.

[17] Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., Chintala, S., 2020. Pytorch distributed: experiences on accelerating data parallel training. Proc. VLDB Endow. 13, 3005–3018. URL: 10.14778/3415478.3415530.

[18] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 318–327. doi:10.1109/TPAMI.2018.2858826.

[19] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167.

[20] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 4768–4777.

[21] Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using deep learning for image-based plant disease detection. Frontiers in Plant Science Volume 7 - 2016. doi:10.3389/fpls.2016.01419.

[22] Power, D.J., 2003. A brief history of decision support systems. DSSResources.COM 4.

[23] Qwen Team, 2025. Qwen3: Think deeper, act faster. https://qwenlm.github.io/blog/qwen3/. Accessed July 15, 2026.

[24] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. p. 1135–1144. doi:10.1145/2939672.2939778.

[25] Roumeliotis, K.I., Sapkota, R., Karkee, M., Tselikas, N.D., 2026. Agentic ai with orchestrator-agent trust: A modular visual classification framework with trust-aware orchestration and rag-based reasoning. IEEE Access 14, 26965–26982. doi:10.1109/ACCESS.2026.3662282.

[26] Roumeliotis, K.I., Sapkota, R., Karkee, M., Tselikas, N.D., Nasiopoulos, D.K., 2025. Plant disease detection through multimodal large language models and convolutional neural networks. URL: https://arxiv.org/abs/2504.20419, arXiv:2504.20419.

[27] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. doi:10.1109/ICCV.2017.74.

[28] Serna, A.M.G., 2025. PlantDoc dataset. https://www.kaggle.com/datasets/andresmgs/plantdec. Kaggle dataset, accessed July 15, 2026.

[29] Shiyan, A.S., Kozlov, I.D., Baimuratov, I.R., Zhukova, N.A., 2025. Recognizing plants and their diseases: Benchmarks for multiclass and multilabel classification. Pattern Recognition and Image Analysis 35, 159–168. doi:10.1134/S1054661825700087.

[30] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N., 2020. Plantdoc: A dataset for visual plant disease detection, in: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, Association for Computing Machinery, New York, NY, USA. p. 249–253. doi:10.1145/3371158.3371196.

[31] Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning (ICML), PMLR. pp. 6105–6114.

[32] Taneja, N., Garg, N., Gupta, S., Kaushal, R., 2023. Comparative analysis of convolutional neural network techniques for plant disease detection, in: 2023 4th International Conference for Emerging Technology (INCET), pp. 1–4. doi:10.1109/INCET57972.2023.10170673.

[33] Wojciuk, M., Swiderska-Chadaj, Z., Siwek, K., Gertych, A., 2024. Improving classification accuracy of fine-tuned cnn models: Impact of hyperparameter optimization. Heliyon 10, e26586. doi:10.1016/j.heliyon.2024.e26586.

[34] Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.J., 2017. Big data in smart farming – a review. Agricultural Systems 153, 69–80. doi:10.1016/j.agsy.2017.01.023.

[35] Xiang, K., Shi, D., Zhu, X., 2026. Quantifying the reliability gap in cross-domain plant disease classification: benchmarking the limited efficacy of standard mitigation techniques under controlled-to-field shift. Frontiers in Plant Science Volume 17 - 2026. doi:10.3389/fpls.2026.1826962.

[36] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E., 2024. A survey on multimodal large language models. National Science Review 11, nwae403. doi:10.1093/nsr/nwae403.