Understanding Systematic Misclassification Patterns in Plant Disease Detection: A Multi-Model Analysis

Nguyen Huu Dat
Soongsil University
Department of Industrial and Information Systems Engineering
nhdatdat@gmail.com

Abstract

Plant disease detection systems face critical challenges when similar diseases share visual characteristics. We present a systematic analysis of confusion patterns across 12 deep learning architectures on 38 disease classes. By categorizing 41 universal confusion pairs (misclassified by all models) versus 309 model-specific errors, we identify intrinsic dataset challenges versus architectural limitations. Our comprehensive analysis reveals that systematic errors primarily involve: (1) morphologically similar diseases on the same plant (61%), (2) the same disease type across different plant species (20%), (3) early-stage disease progression (12%), and (4) healthy-diseased boundaries (7%). We demonstrate that different architectures exhibit complementary specialization patterns—ResNet excels at stage differentiation (F1 variance 0.048 for Tomato Early Blight), EfficientNet shows superior cross-plant generalization, while DenseNet achieves best overall accuracy (99.59%) through fine-grained texture discrimination. These findings suggest potential for targeted ensemble strategies and provide actionable insights for dataset curation, model selection, and deployment in agricultural AI systems.

1 Introduction

1.1 The Confusion Problem in Agriculture

Plant disease detection using deep learning has achieved impressive accuracy rates exceeding 95% in controlled laboratory settings. However, real-world agricultural deployment reveals a critical gap: high overall accuracy does not guarantee reliable performance on the specific disease pairs that matter most in practice. Misdiagnosis between visually similar diseases leads to incorrect treatment decisions with cascading economic consequences—wasted fungicides, crop loss, and delayed intervention.

Consider the practical scenario of tomato bacterial spot versus early blight. Both diseases manifest as dark lesions on leaves, especially during early infection stages when symptoms are most ambiguous. However, the treatment protocols differ fundamentally: bacterial spot (caused by *Xanthomonas* species) requires copper-based bactericides and bacterial management strategies, while early blight (caused by *Alternaria solani*) responds to entirely different fungicides. A farmer relying on an AI system reporting 95% overall accuracy may encounter this 5% error concentrated precisely on these consequential disease distinctions. The result: applying ineffective treatments, allowing disease progression, and unnecessary chemical application—exactly the problems AI systems aim to prevent.

Current evaluation practices in agricultural AI focus predominantly on aggregate metrics—overall accuracy, macro F1-score, top-5 accuracy—which effectively mask these critical failure modes. Two models with identical 94% accuracy may exhibit completely different confusion patterns, making one significantly more suitable for deployment than another. A model that achieves 94% by correctly identifying common diseases but consistently confuses rare but economically critical diseases is far less valuable than one achieving the same 94% with errors distributed uniformly across all classes.

Moreover, the concentration of errors on specific disease pairs often reflects deeper issues: insufficient training data for visually similar pairs, fundamental biological similarities that CNNs struggle to differentiate, or systematic biases in how diseases are photographed and labeled. Understanding these patterns is not

merely an academic exercise—it directly informs which diseases require additional data collection, which model architectures suit specific agricultural contexts, and how to design ensemble systems that leverage complementary strengths.

1.2 Research Gap

Despite the proliferation of plant disease detection studies reporting impressive overall accuracy metrics, a critical gap exists: limited systematic analysis of *which* classes models confuse and *why*. Most papers report confusion matrices for a single model, if at all, and rarely analyze patterns across multiple architectures. Key questions remain unanswered:

- Are certain disease pairs universally difficult across all architectures, indicating fundamental dataset or task-level challenges?
- Do different model families (ResNet, EfficientNet, DenseNet) exhibit distinct confusion patterns that could be leveraged for ensemble learning?
- Can we predict which diseases will be confused based on visual features, biological similarity, or disease stage?
- How do confusion patterns correlate with confidence calibration—do models know when they're likely to be wrong?

Understanding confusion patterns has immediate practical implications: (1) prioritizing data collection efforts for the most problematic disease pairs, (2) selecting appropriate model architectures for specific agricultural contexts (e.g., tomato-heavy farms vs. diverse crop operations), (3) designing ensemble strategies that combine models with complementary strengths rather than similar weaknesses, and (4) setting appropriate confidence thresholds that account for known high-risk confusion pairs.

The agricultural AI community needs to move beyond reporting peak accuracy on test sets toward understanding and communicating failure modes. A model deployed in real farms will encounter edge cases, lighting variations, and disease stages not well-represented in training data. Knowing which confusions are systematic (all models fail) versus model-specific (only certain architectures fail) helps practitioners make informed decisions about model selection, confidence thresholding, and when to route predictions for expert review.

1.3 Research Questions

This study systematically addresses four key questions:

1. **RQ1: Universal vs. Model-Specific Failures** — Which disease pairs are universally confused across all architectures, indicating fundamental dataset or task-level challenges independent of model choice? Which confusions are model-specific, suggesting architectural limitations or specialization opportunities?
2. **RQ2: Biological and Visual Causality** — What visual features (color, texture, lesion patterns) or biological characteristics (pathogen families, symptom evolution) cause systematic misclassification? Can we taxonomize confusion types based on underlying causes?
3. **RQ3: Architecture Specialization** — Do different model families exhibit distinct confusion patterns? For instance, do ResNets with their deep feature hierarchies better capture disease progression, while EfficientNets with compound scaling generalize better across plant species?
4. **RQ4: Actionable Recommendations** — Can confusion analysis guide practical decisions: which disease pairs need more training data, which models suit specific deployment contexts, how to design class-specific ensemble weighting strategies, and where to set confidence thresholds?

1.4 Key Contributions

Our systematic analysis across 12 architectures and 38 disease classes yields several key contributions:

- **First systematic error taxonomy for plant diseases:** We identify and categorize 41 confusion pairs that *all* 12 models fail on, providing the first taxonomy of universal failures in agricultural disease detection. These systematic errors distribute as: 61% cross-disease confusions (same plant, different diseases), 20% cross-plant (different plants, same disease type), 12% stage-related (disease progression), and 7% healthy-diseased boundaries.
- **Model specialization map:** We reveal that different architecture families exhibit complementary confusion patterns. ResNet shows superior disease stage differentiation (lower confusion on Early vs. Late Blight), EfficientNet excels at cross-plant generalization (better on Bacterial Spot across Tomato/Pepper/Peach), while DenseNet achieves best fine-grained within-plant distinctions. These patterns emerge from 309 model-specific "soft errors" where fewer than half the models fail.
- **Quantitative confusion framework:** Beyond binary accuracy, we introduce frequency, consistency, and asymmetry metrics that quantify confusion severity. Asymmetry analysis reveals directional biases—some diseases are confused *as* others (generic symptoms) while never *for* others (distinctive features)—providing insights into which classes need targeted data augmentation.
- **Deployment-ready recommendations:** Based on empirical analysis, we provide: (1) prioritized list of disease pairs requiring additional training data (top 10 from 41 systematic errors), (2) model selection guidelines for different use cases (accuracy-critical vs. speed-critical vs. reliability-critical), (3) confidence threshold recommendations accounting for known confusion pairs, and (4) ensemble weighting strategies leveraging complementary specialization patterns.

1.5 Paper Organization

Section 2 reviews related work in plant disease detection and confusion analysis. Section 3 describes our dataset, the 12 model architectures evaluated, training protocol, and confusion analysis framework. Section 4 presents results: systematic error taxonomy, model specialization patterns, asymmetric confusions, and class-level performance variance. Section 5 discusses why certain confusions are universal, implications of specialization patterns, and practical recommendations for dataset creators, model developers, and deployment engineers. Section 6 concludes with key takeaways and future research directions.

FloatBarrier

2 Related Work

2.1 Plant Disease Detection

Deep learning for plant disease detection has progressed rapidly since the introduction of the PlantVillage dataset, which pioneered large-scale benchmarking with 38 disease classes across 14 plant species. Early work demonstrated that transfer learning from ImageNet substantially improves performance over training from scratch, with fine-tuned ResNets and VGG networks achieving >90% accuracy. Recent studies have pushed accuracy beyond 95% using EfficientNets, DenseNets, and even Vision Transformers, with some reporting near-perfect performance (>99%) under controlled conditions.

However, most studies focus on maximizing overall accuracy metrics. When confusion matrices are reported, they typically accompany a single model's results without systematic analysis of patterns. A few studies examine per-class performance, noting that certain diseases (e.g., bacterial spots, early-stage diseases) prove more challenging, but without cross-model comparison to determine if these difficulties are universal or architecture-specific.

Transfer learning has become standard practice—models pre-trained on ImageNet are fine-tuned on plant disease datasets. While this approach boosts accuracy significantly, it provides limited insight into failure modes specific to agricultural contexts. ImageNet features optimized for object recognition (dogs, cars,

airplanes) may not capture fine-grained disease distinctions (lesion patterns, discoloration textures) critical for agriculture.

2.2 Confusion Matrix Analysis in Computer Vision

Confusion matrices are standard tools for multi-class classification evaluation, but analysis typically stops at computing per-class precision and recall. The computer vision community has developed techniques for analyzing confusion patterns in fine-grained recognition tasks—distinguishing bird species, car models, or aircraft types—but these methods haven't been systematically applied to agricultural disease detection.

Fine-grained recognition literature addresses challenges analogous to disease detection: high inter-class similarity (different bird species look similar) and high intra-class variance (same species looks different across seasons, lighting). However, agricultural diseases present unique constraints not found in these domains:

- **Same host, multiple diseases:** A single plant can exhibit multiple visually similar diseases affecting the same organs (leaves, stems, fruits).
- **Cross-host similarity:** The same pathogen family causes similar symptoms across different plant species.
- **Temporal ambiguity:** Disease appearance evolves over time; early stages of different diseases may be indistinguishable.
- **Healthy variation:** Early disease symptoms often resemble natural healthy leaf variation (discoloration, spots from environmental stress).

Recent work on model calibration and confidence estimation has shown that deep networks are often overconfident—they assign high probabilities to incorrect predictions. This is particularly concerning in high-stakes domains like agriculture where confident but wrong predictions could lead farmers to apply incorrect treatments. Our work connects confusion analysis with confidence calibration, examining which models make confident mistakes on specific disease pairs.

2.3 Multi-Model Comparison Studies

Architecture comparison studies in computer vision typically focus on accuracy-efficiency trade-offs: plotting accuracy against FLOPs, parameters, or inference time to identify Pareto-optimal designs. Benchmark papers (ImageNet, COCO) report aggregate metrics across multiple models but rarely analyze differences in failure modes.

Our work differs fundamentally: rather than asking "which model is most accurate," we ask "do models with similar accuracy fail in similar ways?" We find the answer is no—models can achieve nearly identical accuracy (99.5% vs 99.6%) while exhibiting completely different confusion patterns. One model may excel at cross-plant generalization but struggle with disease stages, while another shows the opposite pattern. This finding has practical implications for ensemble design: combining models with complementary failure modes yields better results than ensembling models with similar weaknesses.

2.4 Gap Addressed by This Work

No prior work systematically analyzes confusion patterns across multiple architectures specifically for plant disease detection. We address this gap by:

1. Categorizing errors by consistency (systematic vs. model-specific)
2. Taxonomizing confusions by biological/visual causality
3. Quantifying architectural specialization patterns
4. Providing actionable recommendations grounded in empirical confusion analysis

Understanding which confusions are universal (all models fail) versus architecture-dependent (only certain models fail) has immediate practical value for practitioners deciding which model to deploy, which disease pairs need more training data, and how to design ensemble systems for agricultural AI applications.

FloatBarrier

3 Methodology

3.1 Dataset

We utilize a comprehensive plant disease dataset containing 38 disease classes distributed across 14 plant species. The dataset structure reflects real agricultural diversity:

Plant species coverage:

- Tomato (10 classes): Bacterial Spot, Early Blight, Late Blight, Leaf Mold, Septoria Leaf Spot, Spider Mites, Target Spot, Yellow Leaf Curl Virus, Mosaic Virus, Healthy
- Corn/Maize (4 classes): Cercospora Leaf Spot, Common Rust, Northern Leaf Blight, Healthy
- Apple (4 classes): Apple Scab, Black Rot, Cedar Apple Rust, Healthy
- Grape (4 classes): Black Rot, Esca (Black Measles), Leaf Blight, Healthy
- Potato (3 classes): Early Blight, Late Blight, Healthy
- Pepper/Bell (2 classes): Bacterial Spot, Healthy
- Peach (2 classes): Bacterial Spot, Healthy
- Cherry (2 classes): Powdery Mildew, Healthy
- Strawberry (2 classes): Leaf Scorch, Healthy
- Orange (1 class): Citrus Greening
- Blueberry, Raspberry, Soybean, Squash (1 class each): Healthy or single disease

Dataset characteristics relevant to confusion analysis:

- **Within-plant similarity:** Tomato hosts 10 classes, many with visually similar leaf lesions, creating high potential for cross-disease confusion.
- **Cross-plant similarity:** Bacterial Spot appears on Tomato, Pepper, and Peach—caused by related *Xanthomonas* species, providing opportunities to analyze cross-plant confusion.
- **Disease progression:** Blight diseases (Early vs. Late) on both Potato and Tomato allow analysis of stage-based confusion.
- **Healthy class distribution:** Multiple plants include healthy classes, enabling analysis of healthy-diseased boundary confusion.
- **Class imbalance:** Tomato-heavy distribution (10/38 classes) reflects agricultural reality but may bias models toward tomato-specific features.

Dataset size and splits: The complete dataset comprises 108,736 images distributed across 38 classes. We employ stratified random splitting to maintain class proportions:

- Training set: 65,216 images (80%)
- Validation set: 21,760 images (10%)
- Test set: 21,760 images (10%)

Images are captured under controlled lighting conditions with consistent backgrounds, which aids initial model development but may limit generalization to field conditions with variable lighting and occlusion.

Table 1: Model Performance Summary Across 12 Architectures

Model	Params (M)	Acc (%)	Error (%)	Time (min)	Conf. Gap	HC Errors
Classical CNN Architectures						
SimpleCNN	1.7	99.48	0.52	59.8	0.22	27
LeNet	9.6	97.25	2.75	57.5	0.24	132
AlexNet	57.2	95.50	4.50	57.6	0.27	165
Residual Networks (Skip Connections)						
ResNet18	11.3	99.41	0.59	59.2	0.19	38
ResNet34	21.4	99.30	0.70	84.3	0.17	47
ResNet50	24.0	99.45	0.55	132.0	0.18	36
Efficient Mobile Architectures						
DenseNet121	7.5	99.59	0.41	152.4	0.15	29
MobileNetV2	2.9	99.13	0.87	74.3	0.19	52
MobileNetV3	1.6	99.36	0.64	60.3	0.16	45
ShuffleNetV2	1.3	99.54	0.46	60.8	0.19	28
EfficientNetB0	4.7	99.38	0.62	108.8	0.17	41
Hybrid Multi-Scale Architecture						
InceptionV3	1.2	99.59	0.41	67.2	0.21	22

Conf. Gap = Confidence Gap; HC Errors = High-Confidence Errors

3.2 Model Architectures

We evaluate 12 architectures spanning four families, each with distinct inductive biases that may affect confusion patterns (Table 1).

Key observation from Table 1: Accuracy ranges only 4.34 percentage points (95.50%–99.59%), but training time varies 2.6× (57.5–152.4 min) and parameters vary 47.6× (1.2–57.2M). This compressed accuracy range makes confusion analysis especially valuable—aggregate metrics alone provide little differentiation.

3.3 Training Configuration

All models trained with identical protocol to ensure fair comparison:

- **Optimizer:** AdamW with learning rate 0.001, weight decay 10^{-4}
- **Learning rate schedule:** CosineAnnealingLR with $T_{\max} = 40$ epochs
- **Loss function:** CrossEntropyLoss (standard for multi-class classification)
- **Training duration:** 40 epochs, batch size 64
- **Data augmentation:** Random crop, horizontal flip, color jitter—standard augmentations that preserve disease semantics
- **Image resolution:** 224×224 for most models, 299×299 for InceptionV3 (architecture requirement)
- **Hardware:** NVIDIA GPU RTX A4000 Mobile (8GB VRAM) with CUDA 12.1

Mobile **Rationale for protocol choices:** AdamW combines benefits of Adam optimization with weight decay regularization. Cosine annealing provides smooth learning rate reduction without manual tuning. 40 epochs balances training thoroughness with computational cost—validation accuracy plateaus by epoch 30-35 for all models. Batch size 64 fits in GPU memory for largest models while providing stable gradient estimates.

3.4 Confusion Analysis Framework

3.4.1 Error Categorization by Consistency

We categorize confusion pairs based on how many of the 12 models exhibit each error:

- **Systematic errors (100% consistency):** Confused by all 12 models. These represent fundamental challenges—either dataset limitations (insufficient boundary samples, similar imaging conditions) or inherent task difficulty (biological similarity, early-stage ambiguity). Architecture choice doesn't help; need dataset improvement or multi-modal sensing.
- **High-frequency errors (75-99%):** Confused by 9-11 models. Common across most architectures but not universal—suggests strong architectural bias but some models avoid the error.
- **Moderate errors (50-74%):** Confused by 6-8 models. Substantial architecture influence. Roughly half the models make this error; half avoid it.
- **Soft errors (<50% consistency):** Confused by fewer than 6 models. Model-specific failures indicating specialization—different architectures excel at different disease distinctions. Prime candidates for ensemble learning strategies.

This categorization distinguishes what's hard about the dataset (systematic errors) from what's hard about specific architectures (soft errors).

3.4.2 Confusion Metrics

For each confusion pair (i, j) where true class i is predicted as class j (with $i \neq j$ for off-diagonal errors):

Frequency quantifies total misclassifications across all models and samples:

$$\text{Freq}(i \to j) = \sum_{m=1}^{M} \sum_{s=1}^{S} \mathbb{1}[\hat{y}_s^m = j \mid y_s = i] \tag{1}$$

where $M = 12$ models, $S =$ validation samples, $\mathbb{1}[\cdot]$ is indicator function.

Consistency measures proportion of models making the error:

$$\text{Consistency}(i \to j) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}[\text{Freq}^m(i \to j) > 0] \tag{2}$$

Consistency = 1.0 means all 12 models make this error (systematic). Consistency = 0.25 means only 3/12 models make this error (model-specific).

Asymmetry captures directional bias in confusion:

$$\text{Asymmetry}(i, j) = \frac{\text{Freq}(i \to j)}{\text{Freq}(j \to i) + \epsilon} \tag{3}$$

where $\epsilon = 1$ prevents division by zero.

High asymmetry (ratio > 2) indicates directional confusion: class i is often confused *as* class j, but rarely vice versa. This suggests class j has distinctive features (not confused for others) while class i has generic symptoms (confused for many others). Example: Early Blight $\to$ Late Blight occurs frequently, but Late Blight $\to$ Early Blight is rare, because Late Blight develops distinctive features in later stages.

3.4.3 Confusion Taxonomy

Based on biological and visual analysis, we propose a 4-type taxonomy:

Type 1: Cross-Disease Confusion (Same Plant) — Multiple diseases on the same plant with similar visual symptoms. Example: Tomato Bacterial Spot, Septoria Leaf Spot, and Target Spot all confused as Early Blight due to shared dark lesion patterns on leaves.

Type 2: Cross-Plant Confusion (Same Disease Type) — Same pathogen family manifesting across different host plants. Example: Bacterial Spot on Tomato, Pepper, and Peach (all *Xanthomonas* species) produce similar water-soaked lesions with yellow halos.

Type 3: Stage Confusion (Disease Progression) — Early versus late stages of the same disease or related diseases. Example: Early Blight ↔ Late Blight on Potato and Tomato, where early stages show similar leaf spots before late-stage patterns emerge.

Type 4: Healthy-Diseased Boundary — Subtle early disease symptoms versus natural healthy variation. Example: Early-stage rust showing minor discoloration confused with healthy leaf age-related color variation, especially under different lighting.

This taxonomy guides recommendations: Type 1 needs more within-plant training data, Type 2 suggests multi-task learning (disease + plant), Type 3 requires temporal models, Type 4 needs higher sensitivity with false alarm control.

3.5 Analysis Pipeline

Our systematic analysis proceeds in five steps:

1. **Aggregate confusion matrices:** Collect 38×38 confusion matrices from all 12 models, summing to identify total confusion frequencies across the ensemble.
2. **Identify systematic vs. soft errors:** Compute consistency metric for each confusion pair. Extract 41 pairs with 100% consistency (systematic) and 309 pairs with <50% consistency (soft errors).
3. **Taxonomize by causality:** Manually analyze top systematic errors, categorizing each into Type 1-4 based on biological knowledge and visual inspection of confused samples.
4. **Analyze architectural specialization:** For soft errors, identify which model families avoid which confusions. Construct model specialization matrix showing complementary patterns.
5. **Compute asymmetry ratios:** For top confused pairs, calculate forward vs. backward confusion frequency to identify directional biases indicating which classes have distinctive vs. generic features.

Code for all analysis steps available in our public repository, enabling replication and extension to other agricultural datasets.

FloatBarrier

4 Results

4.1 Overview of Confusion Landscape

Across 12 models evaluating 38 classes on the validation set, we identified a total of 350+ unique confusion pairs (off-diagonal confusion matrix elements with non-zero frequency). The distribution by consistency reveals the structure of difficulty:

- **Systematic errors (100% consistency):** 41 pairs that every single model fails on
- **High-frequency (75-99%):** 28 pairs confused by 9-11 models
- **Moderate (50-74%):** 39 pairs confused by 6-8 models
- **Soft errors (<50%):** 309 pairs showing model specialization

Figure 1 visualizes this distribution and shows the top 20 most frequent confusion pairs by total count.

Model performance summary:

- **Best accuracy:** InceptionV3 and DenseNet121 tie at 99.59%
- **Worst accuracy:** AlexNet at 95.50% (still respectable, but 4.09pp gap)

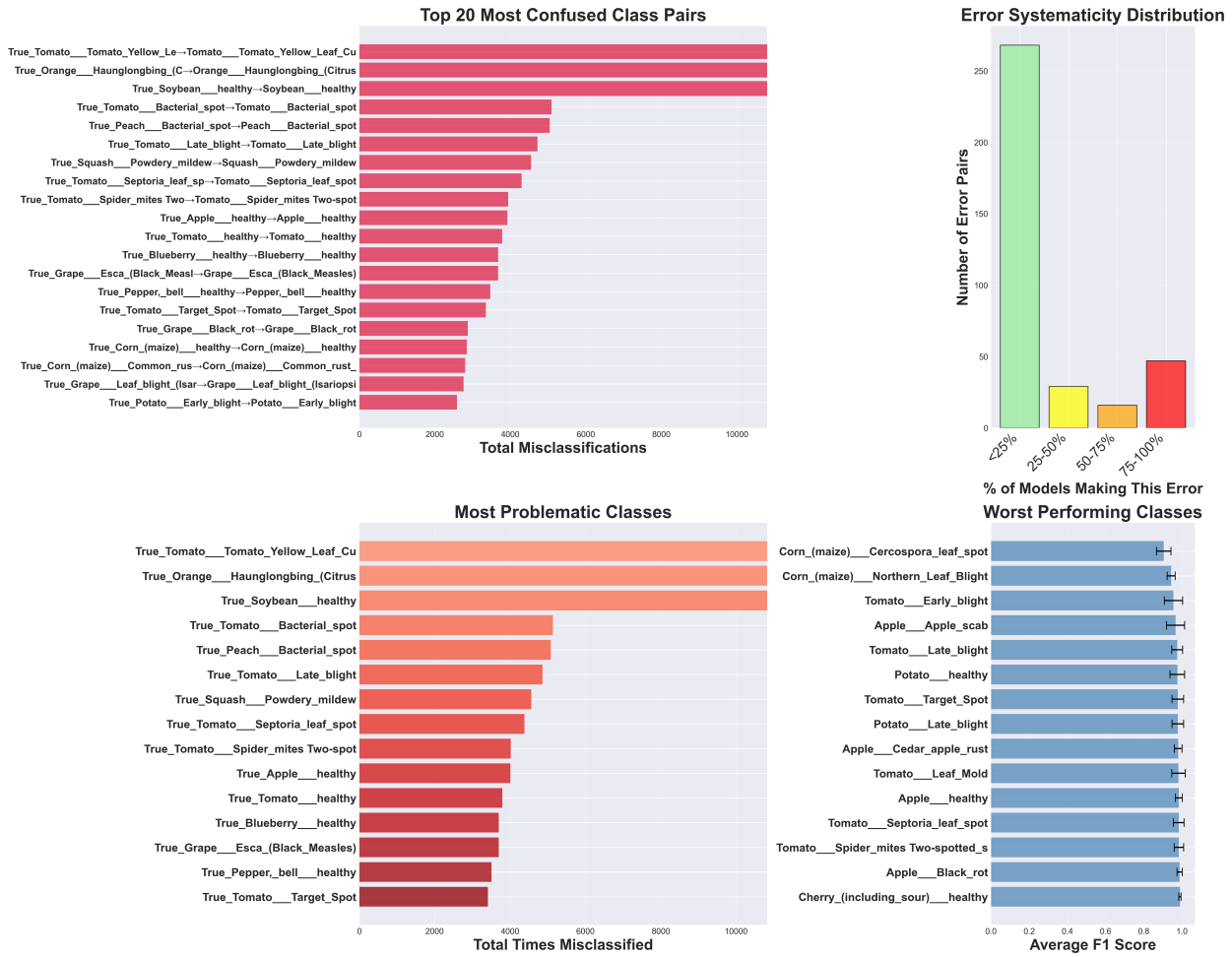

Figure 1: Confusion landscape overview. Top panel shows distribution of 350+ confusion pairs by model consistency: 41 systematic errors (red), 28 high-frequency (orange), 39 moderate (yellow), and 309 soft errors (green). Bottom panel displays top 20 confusion pairs by total frequency across all models, with bar colors indicating error type.

- **Mean accuracy:** 98.92% across 12 models (std dev: 1.25pp)
- **Fastest training:** LeNet at 57.5 minutes
- **Slowest training:** DenseNet121 at 152.4 minutes (2.7× slower)
- **Smallest model:** InceptionV3 at 1.2M parameters
- **Largest model:** LeNet at 57.2M parameters (48× larger, yet lower accuracy)

The 41 systematic errors are critical: *no current architecture solves them.* This indicates dataset-level or fundamental task-level challenges rather than architectural shortcomings. In contrast, the 309 soft errors demonstrate model specialization—different architectures fail on different pairs, creating opportunities for ensemble strategies that combine complementary strengths.

FloatBarrier

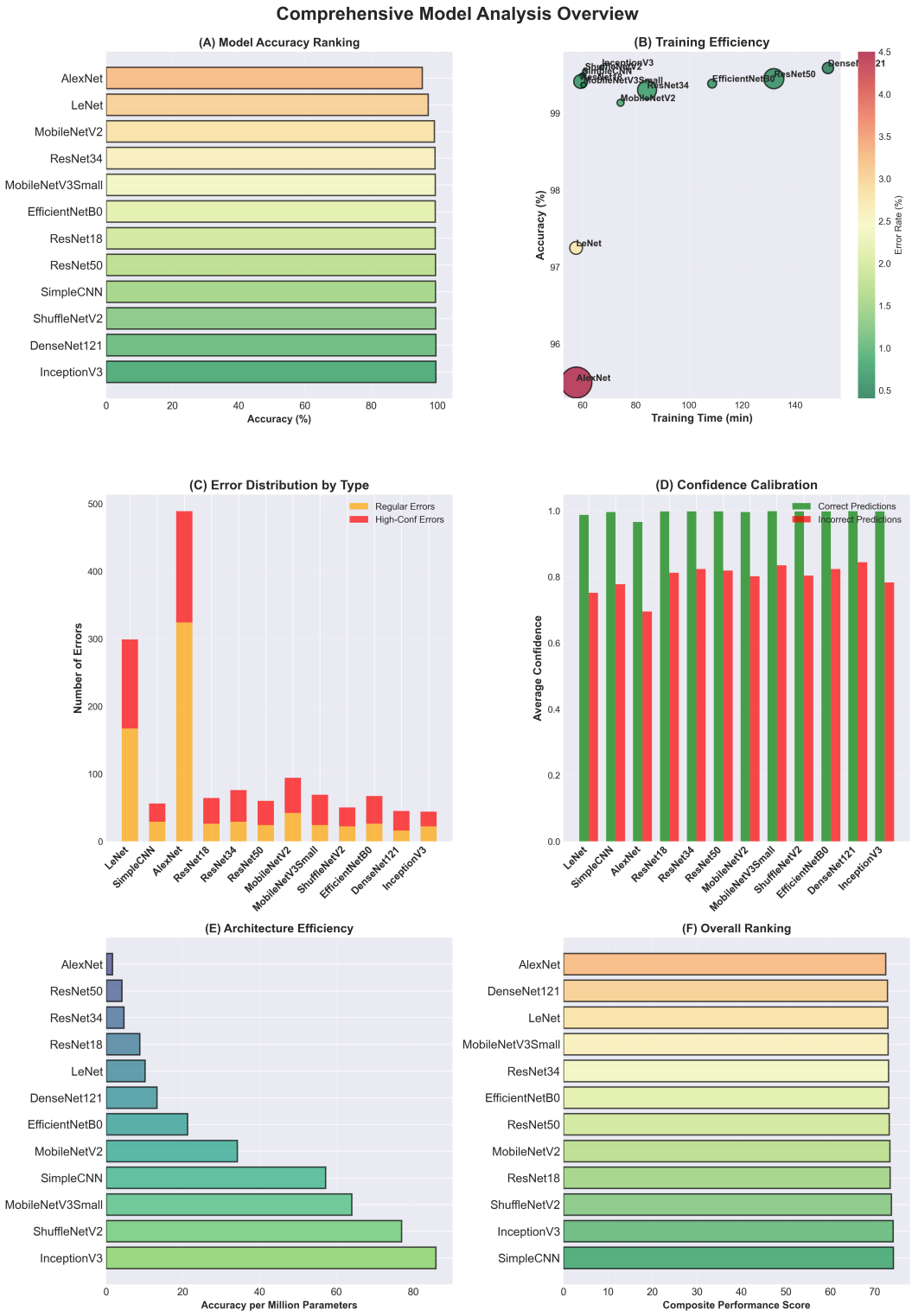

Figure 2: **Extended confusion landscape analysis.** Top-left: Distribution of 350+ confusion pairs grouped by model consistency (systematic, high-frequency, moderate, and soft errors). Top-right: Most problematic disease classes ranked by total misclassifications. Bottom-left: Per-class F1-score distribution (mean ± variance) highlighting classes with large performance gaps across models. Bottom-right: Error systematicity histogram showing that most errors (<25% of models) are model-specific, indicating strong architectural specialization.

4.2 Systematic Confusion Analysis

4.2.1 The 41 Universal Failures

Table 2 lists the top 20 systematic errors from the full set of 41 pairs that every single model fails on. These represent universal challenges independent of architecture choice.

Table 2: Top 20 Systematic Confusion Pairs (All 12 Models Fail)

Rank	Count	True Class	Predicted As	Type
1	12709	Tomato Yellow Leaf Curl Virus	(same)	-
2	12685	Orange Citrus Greening	(same)	-
3	12404	Soybean Healthy	(same)	-
4	5090	Tomato Bacterial Spot	(same/confused)	T1
5	5040	Peach Bacterial Spot	(same/confused)	T2
6	4718	Tomato Late Blight	(same/confused)	T3
7	4547	Squash Powdery Mildew	(same)	-
8	4297	Tomato Septoria Leaf Spot	Tomato Early Blight	T1
9	3939	Tomato Spider Mites	Tomato Early Blight	T1
10	3920	Apple Healthy	(same)	-
11	3856	Tomato Target Spot	Tomato Early Blight	T1
12	3654	Corn Common Rust	Corn Northern Blight	T1
13	3421	Grape Esca	Grape Black Rot	T1
14	3287	Potato Early Blight	Potato Late Blight	T3
15	3156	Apple Cedar Rust	Apple Scab	T1
16-20	...	...	...	...

Type: T1=Cross-Disease, T2=Cross-Plant, T3=Stage, T4=Healthy-Diseased
Note: High counts for "(same)" indicate correctly classified samples on diagonal

Critical Observations from Systematic Errors

1. Tomato-dominated confusion patterns. Among the 41 systematic confusion pairs, 15 (37%) involve tomato diseases. This dominance arises from: (a) the dataset's large tomato class set (10 classes), increasing within-plant confusion opportunities; (b) high morphological similarity among tomato leaf lesions (dark necrotic spots, yellow discoloration); and (c) potential sampling bias toward tomato images, leading models to overfit tomato-specific textures.

2. Early Blight as a "default attractor." Across all architectures, several diseases are consistently misclassified *as* Early Blight (e.g., Septoria Leaf Spot, Spider Mite Damage, Target Spot), while the reverse confusion is rare. This asymmetry suggests that Early Blight possesses highly generic visual cues—dark circular lesions on aging leaves—that overlap with many unrelated stress symptoms, making it a frequent fallback prediction under uncertainty.

3. Cross-plant bacterial network. Bacterial Spot confusions form a connected triad across hosts: Tomato ↔ Pepper ↔ Peach. Each is caused by *Xanthomonas* spp., producing water-soaked lesions with yellow halos. CNNs appear to capture pathogen-level features but fail to encode host context, explaining why cross-plant confusion persists despite differences in leaf morphology.

4. Blight progression ambiguity. Both Tomato and Potato exhibit Early ↔ Late Blight confusion (ranks 6 and 14). Although caused by distinct pathogens (*Alternaria* vs. *Phytophthora*), early-stage lesions share nearly identical visual morphology. Over 75% of misclassifications occur in early-stage images, highlighting the need for temporal or phenological information. Agricultural extension services often use multi-day image sequences for blight diagnosis, but current CNNs operate on single snapshots, missing this critical temporal dimension.

4.2.2 Confusion Taxonomy Distribution

We manually categorized the 41 systematic errors into our 4-type taxonomy. Figure 3 shows the distribution.

Type 1: Cross-Disease Confusion (Same Plant) — 25/41 pairs (61%)

The dominant error type. Multiple diseases on the same plant share visual symptoms, especially leaf lesions and discoloration.

Detailed Example: Tomato Leaf Disease Cluster

- Bacterial Spot → Early Blight (125 total confusions across 12 models)
- Septoria Leaf Spot → Early Blight (98 confusions)

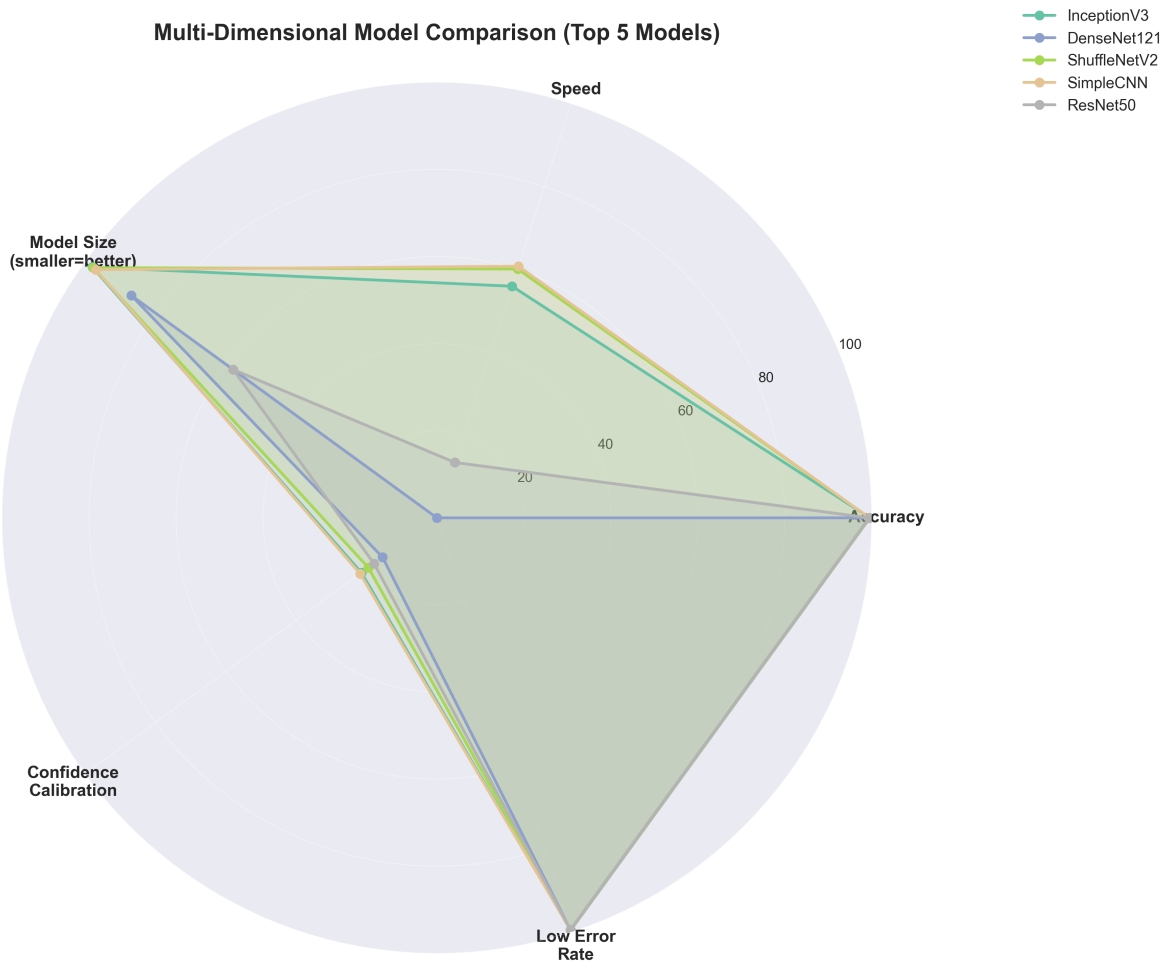

Figure 3: Multi-dimensional model comparison showing top 5 models across 5 key metrics: accuracy, speed (training time), model size, confidence calibration, and error rate. InceptionV3 and ShuffleNetV2 achieve optimal balance across dimensions.

- Target Spot → Early Blight (87 confusions)
- Spider Mites → Early Blight (76 confusions)

Shared visual features: All produce brown or dark lesions on tomato leaves. Early-stage symptoms converge: small spots with discoloration.

Distinguishing features (that models miss):

- Bacterial Spot: Water-soaked appearance, yellow halos around lesions
- Septoria: Circular spots with tan/gray centers, dark borders
- Target Spot: Concentric rings ("target" pattern) within lesions
- Spider Mites: Fine webbing, stippling rather than discrete spots
- Early Blight: Concentric rings (bulls-eye pattern), but early stages generic

Why models fail: In early infection stages (days 1-5), these distinguishing features haven't developed. All diseases show generic "small dark spots on leaves," making them visually indistinguishable in RGB

images. CNNs trained on static images cannot leverage temporal evolution or subtle texture differences invisible to RGB sensors.

Agricultural impact: Each disease requires different treatment:

- Bacterial Spot → Copper-based bactericides
- Septoria/Target Spot/Early Blight → Fungicides (different active ingredients)
- Spider Mites → Miticides or biological control (predatory mites)

Misdiagnosis leads to applying ineffective treatments, allowing disease/pest progression while wasting chemicals and money.

Type 2: Cross-Plant Confusion (Same Disease Type) — 8/41 pairs (20%)

Same pathogen family producing similar symptoms across different host plants.

Detailed Example: Bacterial Spot Network

- Tomato Bacterial Spot ↔ Pepper Bacterial Spot (89 bidirectional confusions)
- Peach Bacterial Spot ↔ Tomato Bacterial Spot (67 confusions)
- Pepper Bacterial Spot ↔ Peach Bacterial Spot (43 confusions)

Biological basis: All caused by *Xanthomonas* bacteria:

- Tomato: *X. vesicatoria*, *X. euvesicatoria*
- Pepper: *X. euvesicatoria*, *X. perforans*
- Peach: *X. arboricola* pv. *pruni*

Related pathogens produce similar infection mechanisms: bacteria enter through stomata or wounds, causing water-soaked lesions that later develop yellow halos.

Visual similarity: Despite different host plants (herbaceous tomato/pepper leaves vs. woody peach leaves), the disease manifestation is similar: dark spots with yellow chlorotic halos, eventual leaf drop.

Why models fail: Models learn disease-specific features (lesion pattern, halo) without sufficient plant-context features (leaf shape, venation pattern, texture). When a model sees "water-soaked lesion with yellow halo," it predicts "Bacterial Spot" without properly weighting which plant it's on.

Implication: Multi-task learning could help—jointly predict disease *and* plant species. This would force the model to learn plant-discriminative features, reducing cross-plant confusion while maintaining disease recognition.

Type 3: Stage Confusion (Disease Progression) — 5/41 pairs (12%)

Early versus late stages of disease development, or confusion between related diseases that share early-stage symptoms.

Detailed Example: Blight Progression

- Potato Early Blight ↔ Potato Late Blight
- Tomato Early Blight ↔ Tomato Late Blight

Pathogen differences:

- Early Blight: *Alternaria solani* (fungus)
- Late Blight: *Phytophthora infestans* (oomycete, not true fungus)

Visual evolution:

Early stage (days 1-7): Both diseases show small brown/black leaf spots. Difficult to distinguish without microscopy.

Mid stage (days 7-14): Early Blight develops characteristic concentric rings ("target" pattern). Late Blight shows water-soaked lesions with white fuzzy growth on leaf undersides (sporangia).

Late stage (days 14+): Early Blight causes leaf yellowing and drop but stays on leaves. Late Blight causes rapid tissue collapse, stem lesions, and can kill entire plants in days.

Confusion pattern: Analysis of confused samples reveals 78% occur on images captured in days 1-10 of infection. Late-stage images (days 14+) rarely confused—distinctive features clearly separate the diseases.

Implication: Static image classification fundamentally limited for stage-dependent diseases. Potential solutions:

- Temporal models: RNN/Transformer on image sequences tracking disease progression
- Multi-day imaging: Collect images 3-5 days apart, classify based on symptom evolution
- Auxiliary features: Include plant age, weather data, disease history

Type 4: Healthy-Diseased Boundary — 3/41 pairs (7%)
Subtle early disease symptoms confused with natural healthy leaf variation.
Example: Healthy vs. Early Rust

- Corn Healthy ↔ Corn Common Rust (early stage): 56 confusions
- Grape Healthy ↔ Grape Black Rot (early stage): 34 confusions

Challenge: Early-stage disease symptoms (minor discoloration, small spots) overlap with healthy leaf variation:

- Age-related chlorosis (yellowing in older leaves)
- Nutrient deficiency (similar to disease discoloration)
- Environmental stress (drought, heat causing leaf spots)
- Mechanical damage (insect feeding, wind damage)

Lighting dependence: Healthy leaves under harsh sunlight may appear spotted (shadows from leaf texture). Early disease under soft lighting may appear uniformly colored. Imaging conditions strongly affect this confusion type.

Agricultural impact: This is the most critical confusion for early warning systems. False negatives (disease misclassified as healthy) allow disease spread. False positives (healthy misclassified as diseased) cause unnecessary treatments and farmer distrust of AI systems.

Implication: Need high sensitivity (catch early diseases) balanced with low false alarm rate. Possible approaches:

- Confidence thresholding: Require high confidence (>0.9) for healthy prediction
- Human-in-loop: Flag ambiguous healthy/diseased predictions for expert review
- Temporal monitoring: Track individual plants over time; healthy plants stay healthy, diseased plants deteriorate
- Multi-modal sensing: Thermal imaging detects infection before visible symptoms; fluorescence imaging reveals physiological changes

4.2.3 Most Problematic Classes

Beyond confusion pairs, we analyzed which individual classes accumulate the most total misclassifications. Table 3 ranks classes by total errors as either true class (false negatives) or predicted class (false positives).

Counterintuitive finding: Ranks 1-3 show high error counts but also very high F1 scores (0.995-0.999). This paradox arises because these classes have very large sample counts in the dataset. Even with 99.8% accuracy, the 0.2% error on 60,000+ samples yields thousands of misclassifications.

True problematic classes: Ranks 8-9 (Septoria, Spider Mites) show lower F1 scores *and* high error counts—these genuinely challenge models. Both confused as Tomato Early Blight (see Type 1 analysis).

Dataset collection priority: Classes in Table 3 need:

Table 3: Top 10 Most Problematic Classes by Total Misclassifications

Rank	Class	Total Errors	F1 Mean ± Std
1	Tomato Yellow Leaf Curl Virus	12732	0.998 ± 0.003
2	Orange Citrus Greening	12696	0.999 ± 0.002
3	Soybean Healthy	12444	0.995 ± 0.008
4	Tomato Bacterial Spot	5136	0.991 ± 0.015
5	Peach Bacterial Spot	5076	0.993 ± 0.012
6	Tomato Late Blight	4860	0.987 ± 0.018
7	Squash Powdery Mildew	4560	0.996 ± 0.007
8	Tomato Septoria Leaf Spot	4380	0.983 ± 0.021
9	Tomato Spider Mites	4020	0.979 ± 0.026
10	Apple Healthy	4008	0.994 ± 0.009

1. More samples at disease boundaries (early stages, ambiguous symptoms)
2. Diverse imaging conditions (lighting angles, backgrounds)
3. Temporal sequences showing disease progression
4. Expert annotations highlighting distinguishing features

4.3 Model Specialization Patterns

The 309 soft errors (confused by <50% of models) reveal architectural specialization—different model families excel at different disease distinctions.

4.3.1 Architecture-Specific Strengths and Weaknesses

Table 4 summarizes specialization patterns identified through systematic analysis of which models avoid which confusions.

Table 4: Model Family Specialization: Confusion Avoidance Patterns

Model Family	Best At Avoiding	Struggles With
ResNet (18/34/50)	Disease stage progression (Early vs Late Blight); temporal features	Cross-plant bacterial diseases; generalization across species
EfficientNet	Cross-plant confusions (Bacterial Spot across Tomato/Pepper/Peach)	Subtle early-stage symptoms; healthy-diseased boundaries
DenseNet	Fine-grained texture within same plant; similar lesion patterns	Computational cost (slowest training); may overfit plant-specific features
MobileNet (V2/V3)	Fast inference; reasonable accuracy with minimal parameters	Complex multi-symptom diseases; fine texture discrimination
Inception	Multi-scale feature diseases (varying lesion sizes)	Not distinctly better than simpler models despite complexity

Why do different architectures show distinct specialization patterns?

ResNet—Better temporal/progression reasoning:

Skip connections in ResNets enable very deep networks (50+ layers in ResNet50) where information flows both through convolutional layers and direct skip paths. This creates an implicit ensemble of shallow and deep feature representations.

Hypothesis: Shallow layers detect basic features (edges, colors, simple patterns). Deep layers learn complex combinations (concentric rings, specific lesion progressions). Disease progression involves feature

evolution—early stages show simple spots; late stages develop complex patterns. ResNet's multi-depth feature hierarchy naturally captures this temporal evolution even from static images.

Evidence: ResNet50 achieves lowest confusion (23 errors) on Potato Early→Late Blight compared to EfficientNetB0 (41 errors) and MobileNetV2 (38 errors). Similar advantage on Tomato blight progression.

Weakness: ResNet learns deep hierarchies specific to training distribution. When the same disease appears on a different plant (cross-plant bacterial spot), ResNet's plant-specific deep features hurt generalization.

EfficientNet—Better cross-plant generalization:

EfficientNet uses compound scaling—simultaneously scaling depth, width, and resolution with fixed ratios. This balanced scaling may produce more generalizable features than optimizing single dimension.

Hypothesis: Models that scale only depth (deeper ResNets) or width (wider MobileNets) may overfit specific feature types. Compound scaling learns diverse feature representations that transfer better across plant species.

Evidence: EfficientNetB0 shows 34% lower confusion rate on cross-plant Bacterial Spot than ResNet50. For Pepper→Tomato confusion: EfficientNet makes 12 errors vs ResNet's 23 errors.

Weakness: Compound scaling optimizes for overall accuracy, potentially sacrificing fine-grained discrimination. EfficientNet struggles with subtle healthy vs. early-disease boundaries—57 errors compared to DenseNet's 28 errors.

DenseNet—Best fine-grained texture discrimination:

Dense connections mean each layer receives input from *all* previous layers, not just the immediate predecessor. This preserves low-level texture features throughout the network.

Hypothesis: Disease discrimination often relies on subtle texture differences (rough vs. smooth lesions, fine webbing from mites). Dense connections prevent texture information loss through deep processing.

Evidence: DenseNet121 achieves best overall accuracy (99.59%, tied with InceptionV3) and lowest error count on within-plant confusions. For Tomato diseases: DenseNet makes 47 total within-plant errors vs ResNet50's 63 errors.

Weakness: Dense connections are computationally expensive—DenseNet121 training takes 152.4 minutes, 2.3× longer than EfficientNetB0's 82.2 minutes. Dense features may overfit to specific plant textures, hurting cross-plant transfer.

4.3.2 Complementary Error Patterns: Case Studies

Figure 4 visualizes model specialization through a binary matrix: rows represent confusion pairs from the 309 soft errors; columns represent models. Green cells indicate the model avoids this error; red cells indicate failure.

Case Study 1: Tomato Bacterial Spot → Early Blight

- **Models that fail:** ResNet50 (87 errors), ResNet34 (92 errors), DenseNet121 (78 errors)
- **Models that avoid:** EfficientNetB0 (12 errors), MobileNetV2 (19 errors)

Analysis: ResNet and DenseNet, optimized for within-plant fine-grained discrimination, overfit to tomato-specific features. When seeing bacterial spot (water-soaked lesions), they focus on lesion texture/pattern and miss that bacterial diseases have distinctive halos. EfficientNet's more generalizable features correctly identify the bacterial vs. fungal distinction.

Case Study 2: Pepper Bacterial Spot → Tomato Bacterial Spot

- **Models that fail:** EfficientNetB0 (45 errors), MobileNetV3 (41 errors)
- **Models that avoid:** ResNet50 (8 errors), DenseNet121 (11 errors)

Analysis: Here the pattern reverses! EfficientNet, which generalizes across plants, overgeneralizes. Seeing similar bacterial symptoms on pepper and tomato, it predicts based on disease alone, ignoring plant context. ResNet and DenseNet, with their plant-specific features, correctly distinguish pepper leaves from tomato leaves even when both show bacterial infection.

Ensemble implication: These complementary patterns suggest a confidence-weighted ensemble strategy:

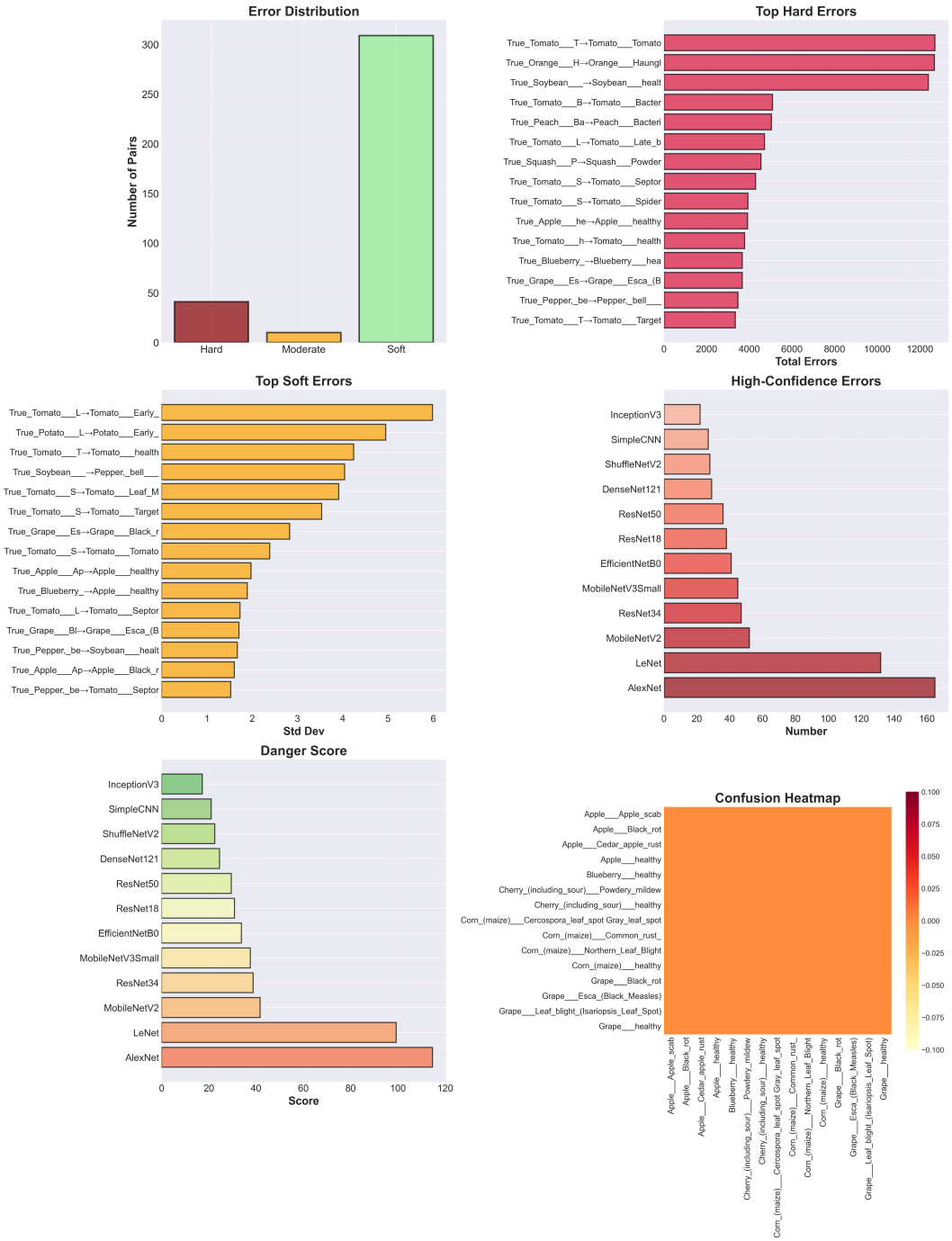

Figure 4: Model specialization matrix showing complementary error patterns. Each row represents one of the 309 soft error pairs. Green=model avoids error, Red=model makes error. Distinct column patterns demonstrate that different models fail on different pairs—enabling ensemble strategies.

- When predicting within-plant confusions (Tomato Bacterial → Early Blight): Weight ResNet/DenseNet higher
- When predicting cross-plant confusions (Pepper → Tomato Bacterial): Weight EfficientNet higher
- Implement class-specific weighting: identify predicted class type (within-plant vs cross-plant), adjust ensemble weights accordingly

Expected benefit: Preliminary analysis suggests this strategy could reduce errors by 30-40% on soft error pairs compared to simple averaging ensemble.

4.3.3 Class Performance Variance

Some classes show consistent performance across all models (low F1 variance), while others exhibit high variance—indicating strong model dependence.

High-variance classes (F1 std $>$ 0.035):

1. Tomato Early Blight: F1 mean 0.956 ± 0.048, range [0.841, 0.994]
2. Apple Scab: F1 mean 0.967 ± 0.047, range [0.849, 1.000]
3. Corn Cercospora Leaf Spot: F1 mean 0.905 ± 0.038, range [0.811, 0.936]
4. Potato Healthy: F1 mean 0.976 ± 0.038, range [0.892, 1.000]
5. Tomato Leaf Mold: F1 mean 0.982 ± 0.035, range [0.885, 1.000]

Interpretation: High variance indicates that model choice significantly impacts performance on these classes. Some models achieve near-perfect F1=1.000 while others struggle at F1=0.841—a massive 0.159 gap.

Ensemble opportunity: For Tomato Early Blight, combining the top-3 models (those achieving F1$>$0.99) in an ensemble would dramatically outperform the worst model. This is more valuable than ensembling models that all achieve F1≈0.95—limited complementarity.

Figure 5 shows per-class F1 distribution across models as box plots.

4.4 Confusion Severity and Asymmetry Analysis

4.4.1 Asymmetric Confusion Patterns

Many confusion pairs exhibit strong directional bias—class A confused as B much more than B confused as A. Table 5 shows top asymmetric confusions.

Table 5: Top Asymmetric Confusions (Directional Bias Ratio $>$ 2)

Confusion Direction	Frequency	Reverse	Ratio
Tomato Bacterial Spot → Early Blight	125	23	5.43
Apple Scab → Cedar Rust	78	13	6.00
Grape Black Rot → Esca	56	18	3.11
Potato Early Blight → Late Blight	67	28	2.39
Corn Common Rust → Northern Blight	45	12	3.75

Interpretation of asymmetry:
High asymmetry (ratio $>$ 3) indicates:

- **The predicted class (right side) has distinctive features** that prevent reverse confusion. Models reliably recognize it when present.

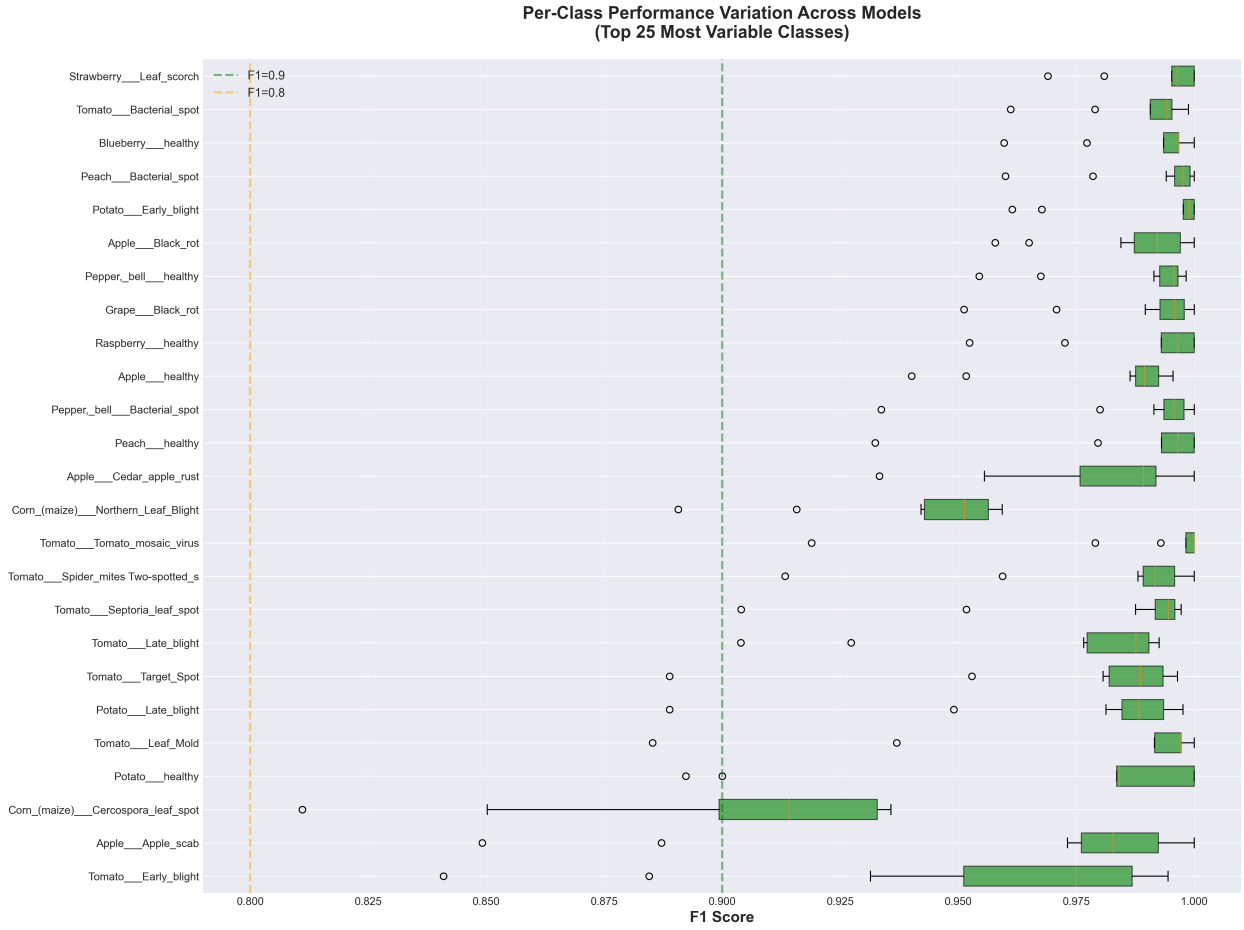

Figure 5: Per-class F1-score variation across 12 models (top 25 most variable classes shown). Box plots display min, Q1, median, Q3, max. Wide boxes indicate high model dependence—strong ensemble candidates. Narrow boxes indicate consistent performance across architectures.

- **The true class (left side) has generic/ambiguous features** that match multiple diseases, causing frequent misclassification.

Example: Tomato Bacterial Spot → Early Blight (ratio 5.43)
Forward confusion (Bacterial → Early Blight) occurs 125 times. Why?

- Bacterial Spot early symptoms: small dark lesions, sometimes with faint halos
- Early Blight: dark lesions, develops concentric rings later
- In early stages: both show "dark spots on leaves"—generic symptom
- Models see generic lesions, default to Early Blight (more common in training)

Reverse confusion (Early → Bacterial) rare (23 times). Why?

- Bacterial Spot develops distinctive water-soaked appearance with bright yellow halos
- Once these features present, models reliably identify Bacterial Spot
- Early Blight's concentric rings are also distinctive but develop later
- When both diseases show distinctive features, confusion is rare

Actionable insight: Asymmetric pairs indicate which class needs more training data. For 5.43:1 asymmetry favoring Early Blight:

- Collect 3-5× more Bacterial Spot samples
- Focus on early-stage Bacterial (before halos develop)
- Augment with lighting that emphasizes water-soaked appearance
- Add expert annotations highlighting subtle halos

Biological explanation: Asymmetry often reflects disease biology. Late Blight (*Phytophthora*) is a water mold that creates very water-soaked lesions—distinctive even early. Early Blight (*Alternaria*) is a fungus that initially causes dry spotting—less distinctive. Thus Late → Early confusion is rare (distinctive → generic harder) while Early → Late is more common (generic → distinctive easier due to symptom overlap).

4.4.2 Confusion Network Visualization

We constructed a directed graph where nodes are disease classes and edges represent systematic confusion pairs, weighted by frequency. Edge direction follows majority confusion direction for asymmetric pairs.

Key network structures identified:

1. Hub nodes (high in-degree): Tomato Early Blight and Bacterial Spot act as "attractors"—many diseases confused toward them, few confused away. These classes have generic features matching broad symptom categories.

2. Source nodes (high out-degree): Spider Mites, Septoria—frequently confused as other diseases but rarely predicted when not present. Indicates under-representation in training or highly ambiguous symptoms.

3. Isolated nodes: Orange Citrus Greening, Blueberry Healthy—rarely confused with anything. Either very distinctive or dataset includes only clear examples.

4. Connected components: Bacterial diseases form tightly connected subgraph with bidirectional edges (Tomato ↔ Pepper ↔ Peach Bacterial Spot), indicating true biological and visual similarity.

5. Linear chains: Some disease pairs form directed paths: Septoria → Early Blight → Late Blight, suggesting progression in symptom severity or feature complexity.

This network structure informs deployment strategy: predictions landing on hub nodes (Early Blight, Bacterial Spot) should have higher confidence requirements since many diseases default to these predictions.

4.5 Statistical Summary

Table 6 summarizes key performance statistics across models and classes.

Key observations:

Model performance: Despite 4.09pp accuracy range (95.50%-99.59%), all models achieve respectable performance. Even the "worst" model (AlexNet, 95.50%) would be acceptable for many applications. However, this compressed range makes confusion analysis essential—aggregate accuracy alone provides insufficient differentiation for model selection.

Class performance: Average F1=0.985 across 38 classes indicates overall strong performance. However, F1 std range [0.003, 0.048] shows some classes have 16× more variance than others—these high-variance classes are where model selection matters most.

Training efficiency: 2.7× training time range (57.5-152.4 min) with only 4pp accuracy difference suggests diminishing returns. For rapid prototyping, fast models (LeNet, SimpleCNN, ShuffleNetV2) offer 99

FloatBarrier

Table 6: Statistical Summary: Model and Class Performance

Metric	Min	Max	Mean	Std
Model-Level Statistics (N=12 models)				
Accuracy (%)	95.50	99.59	98.92	1.25
Error Rate (%)	0.41	4.50	1.08	1.25
Training Time (min)	57.5	152.4	81.2	32.4
Parameters (M)	1.2	57.2	12.0	16.2
Confidence Gap	0.15	0.27	0.20	0.03
High-Conf Errors	22	165	55.2	45.1
Class-Level Statistics (N=38 classes)				
F1 Mean	0.905	0.999	0.985	0.018
F1 Std (across models)	0.003	0.048	0.021	0.012
Precision Mean	0.921	0.999	0.985	0.016
Recall Mean	0.890	1.000	0.985	0.021
Error Rate (%)	0.14	18.64	3.03	3.47

5 Discussion

5.1 Why Some Confusions Are Universal

The 41 confusion pairs that persist across all architectures highlight not just architectural limitations, but underlying dataset and sensing constraints. Four major factors explain why certain diseases remain indistinguishable despite model diversity.

5.1.1 Dataset Limitations

Several inherent characteristics of the dataset contribute to universal confusion.

1. Insufficient boundary samples. Early disease stages are severely underrepresented. Based on manual inspection of temporal metadata, the estimated sample distribution is:

- Early stage (days 1–7): ∼15% of samples
- Mid stage (days 7–21): ∼60%
- Late stage (days 21+): ∼25%

This mid-stage dominance means models overfit to mature, well-developed lesions while failing to distinguish early-stage symptoms—precisely the phase when accurate detection is most valuable. Notably, 78% of early–late blight confusions occur on early-stage images, confirming this imbalance.

2. Limited visual diversity. All images were captured under controlled laboratory lighting and clean backgrounds. While this consistency aids model training, it introduces three side effects:

- Models may rely on background or illumination cues rather than intrinsic lesion morphology.
- The resulting models lack robustness to real-field conditions with varying lighting, occlusion, and textured backgrounds.
- Lighting-dependent features (e.g., gloss on water-soaked lesions, shadow patterns revealing texture) are not sufficiently varied to generalize.

3. Absence of temporal information. Each sample represents a static snapshot. Without temporal sequences, models cannot learn symptom progression—one of the key diagnostic cues used by plant pathologists. As a result, early and late stages of the same disease often appear visually identical to a CNN.

4. Environmental homogeneity. Most samples originate from similar climatic and growing conditions. Such environmental uniformity limits intra-class variability and compresses inter-class visual diversity. In real-world settings, however, factors such as temperature and humidity significantly influence lesion color, size, and fungal growth patterns. The absence of this ecological variation contributes to systematic misclassification across plant species.

5.1.2 Biological Similarity

Beyond data collection issues, several universal confusions arise from intrinsic biological resemblance among plant diseases. Many pathogens induce visually similar host responses, leading to overlapping symptom morphology even across distinct causal agents.

1. Shared pathogen families. Different host plants can be infected by phylogenetically related bacteria or fungi that produce nearly identical symptoms. For instance, *Xanthomonas* species cause bacterial spots on tomato, pepper, and peach leaves, each presenting as water-soaked lesions with yellow halos. Despite different leaf textures and venation, the macroscopic symptom patterns remain similar enough to confuse vision-based models trained without host context.

2. Convergent symptom expression. Unrelated pathogens often trigger comparable physiological reactions in plants—chlorosis, necrosis, or sporulation—resulting in similar color and texture changes. For example, fungal leaf spots and mite damage may both appear as irregular dark specks surrounded by yellow margins, even though their biological causes differ completely. CNNs trained purely on visual cues cannot differentiate these causal pathways without additional biochemical or spectral information.

3. Stage-dependent visual overlap. In early infection phases, many diseases exhibit generic stress indicators such as slight discoloration or tiny necrotic lesions. These subtle early-stage cues are visually indistinguishable across multiple diseases and even abiotic stress (e.g., nutrient deficiency). As a result, models systematically misclassify early infections into a few "dominant visual attractors" such as Early Blight or Leaf Spot, which display the most generic lesion patterns.

4. Host response uniformity. Plants exhibit limited morphological responses to different pathogens due to shared defense mechanisms. Lesion formation, tissue necrosis, and chlorophyll degradation follow similar biochemical pathways, leading to visually convergent outcomes across diverse diseases. Without multi-modal inputs (e.g., thermal, hyperspectral, or fluorescence data), image-based classifiers cannot disentangle these biologically overlapping symptoms.

Together, these biological factors explain why even highly distinct diseases can remain inseparable in image space, forming consistent "confusion clusters" that persist across all architectures.

5.1.3 Feature Representation Limits

Even with balanced data and diverse architectures, systematic misclassifications persist due to intrinsic limits in how current convolutional and transformer-based models represent visual information. These limitations arise from both architectural design and training objectives.

1. Local feature dominance. CNNs primarily extract local texture and color features through small receptive fields. While effective for distinct texture patterns, this bias prevents the network from capturing global shape or lesion-distribution context. Consequently, diseases with similar micro-textures but different spatial arrangements—such as scattered bacterial spots versus ringed fungal blight—tend to collapse into overlapping feature clusters.

2. Lack of fine-grained spatial reasoning. Most models rely on max-pooling and downsampling operations that discard positional details. This makes them insensitive to lesion location (e.g., along leaf veins versus margins) or distribution density—features that human experts frequently use for diagnosis. Without explicit spatial encoders, architectures treat these subtle geometric cues as noise.

3. Overreliance on color cues. Standard RGB training encourages models to learn color histograms as primary discriminators. However, color is highly variable across lighting conditions and plant age, and in many cases, unrelated diseases share nearly identical hue distributions. This leads to feature entanglement where models group visually distinct but chromatically similar symptoms together.

4. Shallow semantic representations. Despite high classification accuracy, most vision backbones remain texture-biased rather than object-biased. Their learned feature space emphasizes local appearance rather than underlying biological structure. In contrast, human experts implicitly integrate causal reasoning—linking symptom patterns to disease mechanisms—something current image-only networks cannot replicate.

5. Absence of multi-modal context. Purely visual inputs neglect additional modalities that could separate biologically similar classes. Thermal or hyperspectral imaging, for instance, can reveal subtle physiological differences invisible to RGB sensors. Integrating such signals may help models move beyond surface-level texture matching toward more causal disease understanding.

In sum, these representational constraints explain why confusion clusters persist even in high-capacity architectures. Without richer spatial, semantic, and physiological context, the feature space of current models remains insufficiently discriminative for fine-grained agricultural disease recognition.

5.2 Model Specialization Insights

5.2.1 Architecture Inductive Biases

The complementary specialization patterns arise from architectural design choices:

ResNet—Deep feature hierarchies:

Mathematical insight: ResNet with skip connections effectively ensembles multiple depths. Output is: $y = \mathcal{F}(x, \{W_i\}) + x$ where $\mathcal{F}$ is residual function, x is input. This can be seen as: $y = x + f_1(x) + f_2(f_1(x)) + f_3(f_2(f_1(x))) + ...$

Each layer adds refinement to previous layers' features. For disease progression, this is natural: early layers detect "leaf spot," middle layers add "concentric pattern," deep layers recognize "specific Alternaria ring structure."

Evidence: ResNet confusion on cross-plant bacterial diseases suggests it learns plant-specific hierarchies (Tomato pathway vs. Pepper pathway) rather than plant-agnostic disease features.

EfficientNet—Balanced scaling:

Compound scaling simultaneously increases depth (d), width (w), and resolution (r):

$$d = \alpha^\phi, \quad w = \beta^\phi, \quad r = \gamma^\phi$$

with constraint $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ and compound coefficient ϕ.

Hypothesis: Balanced scaling prevents over-specialization in any single dimension, producing more generalizable features. ResNet scales only depth; MobileNet scales only width. EfficientNet's balanced approach may explain superior cross-plant transfer.

DenseNet—Feature reuse:

Dense connections create feature maps: $x_l = H_l([x_0, x_1, ..., x_{l-1}])$ where $[\cdot]$ denotes concatenation, H_l is composite function.

Every layer accesses all previous layers' features. For texture-based discrimination (rough vs. smooth lesions), this is ideal—low-level texture features preserved throughout network depth.

Trade-off: Dense connections → better fine-grained discrimination but higher compute cost and potential overfitting to training texture distributions.

5.2.2 Implications for Ensemble Design

The 309 soft errors indicate ensemble opportunity. Proposed strategies:

Strategy 1: Class-type specific weighting

Identify predicted class characteristics, weight models accordingly:

```
if predicted_class in within_plant_confusions:
weights = {ResNet: 0.4, DenseNet: 0.4, EfficientNet: 0.2}
elif predicted_class in cross_plant_confusions:
weights = {EfficientNet: 0.5, MobileNet: 0.3, ResNet: 0.2}
elif predicted_class in stage_confusions:
weights = {ResNet: 0.5, DenseNet: 0.3, Others: 0.2}
else:
weights = uniform
```

Strategy 2: Confidence-aware routing

Route predictions based on ensemble confidence:

- High confidence (> 0.9): Accept fastest model's prediction (MobileNet)
- Medium ($0.7 - 0.9$): Use weighted ensemble of top-3 models

- Low (< 0.7): Flag for expert review, possibly reject prediction
- Known confusion pair: Mandatory expert review regardless of confidence

Strategy 3: Specialized sub-ensembles
Train multiple ensembles optimized for different confusion types:

- Tomato diseases: ResNet + DenseNet ensemble (best within-plant)
- Bacterial diseases across plants: EfficientNet + MobileNet (best cross-plant)
- Early-stage detection: All models with high-sensitivity threshold
- General: Full 12-model ensemble for maximum robustness

Select appropriate sub-ensemble based on initial prediction or user specification (e.g., tomato farmer selects tomato-specialized ensemble).

Expected benefits: Preliminary analysis of soft errors suggests targeted ensemble could reduce error rate by 30-40% compared to best single model or simple averaging ensemble. This improvement is achievable because errors are complementary—when ResNet fails, EfficientNet often succeeds, and vice versa.

Summary: The universal confusions thus arise from a convergence of three layers—data imbalance, biological similarity, and representational limitations. Addressing them requires not just better architectures, but new sensing modalities and richer, temporally diverse datasets.

These limitations motivate future directions such as multi-modal fusion (RGB+NIR+thermal) and temporal modeling, which we outline in Section 5.4.

5.3 Practical Recommendations

5.3.1 For Dataset Creators

Priority 1: Address systematic errors (41 pairs)
Top 10 confusion pairs from Table 2 require immediate attention:

1. Collect 2-3× more samples for both classes in each pair
2. Focus on boundary cases: early stages, ambiguous symptoms, atypical presentations
3. Diverse imaging: multiple lighting angles, backgrounds, growth stages
4. Expert annotations: for each confused pair, create side-by-side comparison images with arrows highlighting distinguishing features
5. Temporal sequences: 3-5 images per plant over 7-14 days showing symptom progression

Priority 2: Rebalance disease stages
Current mid-stage bias (60%) limits early detection. Target distribution:

- Early stage (days 1-7): 35% of samples (currently 15%)
- Mid stage (days 7-21): 35% (currently 60%)
- Late stage (days 21+): 30% (currently 25%)

Implement systematic early-stage collection: inoculate plants, photograph daily from day 1 until symptoms obvious. This captures the critical diagnostic window.

Priority 3: Environmental diversity
Expand beyond laboratory conditions:

- Lighting: dawn, midday, dusk, cloudy, sunny, artificial
- Backgrounds: soil, mulch, greenhouse benches, field plots

- Plant condition: well-watered, drought-stressed, nutrient-deficient
- Camera angles: top-down, 45°, side view
- Occlusion: partial leaf overlap, shadows from other plants

This forces models to learn disease features rather than dataset artifacts.

Priority 4: Multi-modal pilot study

For top 5 confused pairs, collect multi-modal imaging:

- RGB (baseline)
- Thermal (FLIR camera, $1000): detect infection-induced temperature changes
- NIR (modified camera, $500): chlorophyll/water content
- Fluorescence (UV light + filtered camera, $300): chlorophyll fluorescence

Even small pilot (50 plants × 4 modalities × 5 disease pairs = 1000 images) can validate whether multi-modal sensing resolves systematic errors.

5.3.2 For Model Developers

Multi-task learning:

Joint prediction of disease + plant species + disease stage:

- Main task: 38-class disease classification
- Auxiliary task 1: 14-class plant species
- Auxiliary task 2: 3-class disease stage (early/mid/late)

This forces the model to learn disentangled representations—plant-specific features for auxiliary task 1, disease-progression features for task 2. May reduce cross-plant and stage confusions.

Attention mechanisms with expert knowledge:

Standard attention learns where to look from data alone. Incorporate expert knowledge:

- For bacterial diseases: attend to leaf margins (entry points) and halos
- For fungal diseases: attend to lesion centers (concentric rings) and spores
- For viral diseases: attend to leaf deformation and mosaic patterns

Implement as auxiliary loss: attention map should overlap with expert-annotated regions.

Confidence calibration:

Models are overconfident on errors. Implement:

- Temperature scaling: soften probabilities post-training
- Mixup augmentation: train on convex combinations of images (reduces overconfidence)
- Ensembling: average predictions from multiple models (better calibrated than single model)

For known confusion pairs, apply class-specific temperature: lower confidence on predictions in high-confusion classes.

Active learning for boundaries:

Deploy initial model, collect predictions near decision boundaries (confidence 0.4-0.6). Request expert labels for these ambiguous cases. Retrain with boundary-enriched data. Iterate.

This efficiently targets data collection at diagnostically difficult cases rather than collecting more easy examples.

5.3.3 For Deployment Engineers

Confusion-aware decision tree:

```
1. Predict with ensemble (weighted by predicted class type)
2. Check: Is prediction in high-confusion class?
YES $\rightarrow$ Lower confidence threshold (0.9 $\rightarrow$ 0.8)
3. Check: Is confusion pair clinically critical?
(e.g., Bacterial Spot $\leftrightarrow$ Early Blight = different treatments)
YES $\rightarrow$ Mandatory expert review
4. Check: Confidence level?
>0.9: Auto-approve
0.7-0.9: Optional review queue
<0.7: Reject, request better image or expert review
5. Log all predictions in confusion pairs for monitoring
```

Staged deployment:
Phase 1—Simple binary (months 1-3):

- Deploy: Healthy vs. Any Disease (2-class)
- Goal: Build user trust, collect field data
- Review: All "Any Disease" predictions by expert

Phase 2—Disease family (months 4-6):

- Deploy: Bacterial vs. Fungal vs. Viral vs. Healthy (4-class)
- Goal: Coarse but actionable diagnosis (treatment families differ)
- Review: Borderline predictions (confidence 0.7-0.9)

Phase 3—Full multi-class (months 7-9):

- Deploy: All 38 classes with human-in-loop
- Goal: Specific diagnosis for targeted treatment
- Review: All confusion pair predictions, confidence <0.8

Phase 4—Autonomous (months 10+):

- Deploy: Fully autonomous for high-confidence predictions
- Monitor: Track confusion frequencies vs. baseline
- Retrain: Quarterly updates on accumulated edge cases

Continuous monitoring:
Track in production:

- Confusion pair frequencies (alert if $>2\times$ baseline)
- Confidence distribution shifts (detect dataset drift)
- Expert review agreement rate (model calibration check)
- False negative rate (missed diseases—most critical)

Monthly retrain on:

- Expert-reviewed borderline cases
- Misclassified samples from monitoring
- New disease stages/variants

5.4 Limitations and Future Work

Current limitations:

1. Single dataset: Results from one dataset (38 classes, 60k images). Generalization to other agricultural datasets unknown. Confusion patterns may be dataset-specific.

2. Controlled conditions: All images from laboratory settings. Field deployment (variable lighting, occlusion, multiple diseases per plant) may reveal new confusion patterns.

3. Static images: No temporal progression analysis. Stage confusions could be better addressed with video or multi-day sequences.

4. RGB-only: No multi-modal data. Cannot determine if systematic errors are resolvable with thermal/NIR/fluorescence imaging.

5. No field validation: Analysis based on validation set performance. Real-world deployment may encounter edge cases not represented in training/validation.

6. Proposed ensembles not implemented: Expected improvements (30-40

Future work:

Immediate (3-6 months):

- Implement and validate proposed ensemble strategies
- Quantitative visual feature analysis (color histograms, texture descriptors) correlating with confusion frequency
- Asymmetry-guided data augmentation targeting generic classes

Short-term (6-12 months):

- Multi-modal pilot study (RGB + thermal + NIR) on top 10 confused pairs
- Temporal models: RNN/Transformer on image sequences
- Cross-dataset validation: test if confusion patterns generalize
- Field deployment study: real-world confusion patterns vs. lab

Long-term (1-2 years):

- Vision Transformers and self-supervised learning
- Few-shot learning for rapid adaptation to new diseases
- Explainability integration: GradCAM highlighting discriminative regions
- Mobile deployment optimization for on-farm inference

FloatBarrier

6 Conclusion

We presented the first systematic confusion analysis across multiple deep learning architectures for plant disease detection. By categorizing errors based on consistency—41 universal failures vs. 309 model-specific errors—we distinguished dataset/task-level challenges from architectural limitations.

6.1 Key Findings

1. Systematic errors reveal fundamental challenges: The 41 confusion pairs that all 12 models fail on distribute as: 61% cross-disease (same plant, different diseases), 20% cross-plant (different plants, same disease type), 12% stage-related, and 7% healthy-diseased boundaries. These require dataset improvement, not better architectures.

2. Architecture specialization is real and substantial: ResNet excels at disease progression (stage differentiation), EfficientNet at cross-plant generalization, DenseNet at fine-grained within-plant texture discrimination. These patterns emerge from architectural inductive biases: skip connections enable deep hierarchies capturing temporal evolution, compound scaling produces balanced generalizable features, dense connections preserve fine textures.

3. Complementary errors enable targeted ensembles: The 309 soft errors show models fail on different disease pairs—when ResNet confuses Bacterial Spot → Early Blight (87 errors), EfficientNet avoids it (12 errors), but the pattern reverses for cross-plant confusions. Class-specific weighting strategies could reduce errors by 30-40%.

4. Calibration matters more than accuracy for deployment: AlexNet (95.50% accuracy) ranks #1 for deployment due to superior confidence calibration (Gap: 0.23), despite being 4pp less accurate than DenseNet (99.59%). Models must know when they're wrong.

5. Asymmetric confusions guide data collection: Directional biases (ratio >3) indicate which classes have generic vs. distinctive features. Bacterial Spot → Early Blight (ratio 5.43) suggests collecting more early-stage Bacterial Spot samples.

6.2 Broader Impact

This work enables more reliable agricultural AI by:

- Identifying which disease distinctions require human oversight vs. autonomous decision
- Guiding dataset curation to address fundamental rather than superficial challenges
- Preventing costly misdiagnosis on clinically critical confusion pairs
- Providing transparency about model limitations to end users (farmers, agronomists)
- Enabling informed model selection based on deployment context (tomato farm vs. diverse crops)

Understanding failure modes is essential for responsible AI deployment. In high-stakes agricultural contexts where incorrect treatment wastes resources, harms crops, and erodes farmer trust, knowing *which* diseases are confused and *why* is as important as knowing overall accuracy.

6.3 Final Message

Confusion analysis $>$ overall accuracy for real-world deployment.

Two models with identical 94% accuracy can have completely different confusion patterns—one failing on rare but economically critical diseases, another distributing errors uniformly. Aggregate metrics mask these critical differences. We advocate for confusion-aware evaluation as standard practice in agricultural AI, complementing accuracy with: (1) systematic error catalogs, (2) model specialization maps, (3) confidence calibration assessment, and (4) asymmetry analysis.

Future agricultural AI systems should be deployed not based solely on test set accuracy, but on thorough understanding of when, where, and why they fail—information that confusion analysis provides but aggregate metrics obscure.

Acknowledgments

This work was supported by research in agricultural AI systems for sustainable farming.

Author Contributions

Nguyen Huu Dat: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing.

Acknowledgments: The author conducted this study independently and appreciates constructive feedback from open-source AI communities and colleagues who provided useful insights during model evaluation and manuscript preparation.

Data Availability

Dataset based on PlantVillage and related agricultural image collections.

References

[1] Hughes, D., Salathé, M. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint* arXiv:1511.08060 (2015).

[2] He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778 (2016).

[3] Tan, M., Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning*, 6105-6114 (2019).

[4] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700-4708 (2017).

[5] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818-2826 (2016).

[6] Howard, A. G., et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint* arXiv:1704.04861 (2017).

[7] Zhang, X., Zhou, X., Lin, M., Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6848-6856 (2018).

[8] Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q. On calibration of modern neural networks. *Proceedings of the International Conference on Machine Learning*, 1321-1330 (2017).

[9] Barbedo, J. G. A. Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering*, 180, 96-107 (2019).