

Benchmarking Vision Transformers Against CNNs for Plant Disease Diagnosis: Performance, Robustness, and Visual Interpretability

Vatsalya Betala

B.Sc. (Hons) Computer Science and Mathematics
Krea University

Supervised by

Prof. Chiranjib Sur
Krea University

Abstract

Plant disease diagnosis from leaf imagery has traditionally relied on convolutional neural networks (CNNs), which achieve strong performance through localized feature extraction. Recent advances in Vision Transformers (ViTs), which model global spatial relationships via self-attention, suggest potential advantages for capturing distributed disease patterns. However, whether improved predictive performance corresponds to meaningful differences in underlying feature representations remains insufficiently understood.

In this work, we present a controlled comparison of CNN and Vision Transformer architectures within a modular plant disease diagnosis pipeline that decomposes the task into three sequential stages: leaf localization, species classification, and disease prediction. This modular design enables isolation of classification behavior by standardizing input structure and reducing background-induced variability. Empirically, Vision Transformers achieve higher classification accuracy than CNN-based models under identical training conditions, while exhibiting increased sensitivity to optimization hyperparameters.

Through comparative interpretability analysis using Grad-CAM and attention visualizations, we observe a consistent divergence in feature attribution: CNN models predominantly focus on localized lesion regions, whereas Vision Transformers distribute attention across broader leaf-level structures. These findings suggest that performance improvements in Vision Transformers are accompanied by a qualitative shift in representational behavior, moving from localized texture dependence toward more globally contextual reasoning. This work highlights the role of architectural inductive bias in shaping both predictive performance and feature utilization, and underscores the importance of controlled evaluation frameworks for understanding model behavior beyond aggregate accuracy metrics.

Keywords: plant disease classification, convolutional neural networks, Vision Transformers, inductive bias, Grad-CAM, interpretability, modular pipeline

Contents

1	Introduction	4
2	Conceptual Foundations	5
2.1	Inductive Bias in Learning Systems	5
2.2	Inductive Bias in Convolutional Neural Networks	5
2.3	Inductive Bias in Vision Transformers	6
2.4	Representation, Feature Utilization, and Interpretability	6
2.5	Linking Inductive Bias to Observed Behavior	7
3	Problem Formulation	7
3.1	Task Definition	7
3.2	Modular Decomposition	8
3.3	Controlled Architectural Comparison	8
3.4	Training Objective	8
3.5	Evaluation Objectives	9
3.6	Scope and Assumptions	9
4	Methodology	9
4.1	Dataset	9
4.2	Modular Pipeline Implementation	10
4.3	Model Architectures	11
4.4	Training Configuration	12
4.5	Evaluation Metrics	14
5	Experiments and Results	14
5.1	Leaf Localization Performance	14
5.2	Classification Performance (RQ1)	14
5.3	Optimization Behavior and Stability	15
5.4	Inference Efficiency	15
5.5	Representational Analysis (RQ2)	16
5.5.1	Saliency Entropy	16
5.5.2	Top- k Mass Concentration	17
5.5.3	Lesion Region Overlap	17

5.5.4	Summary of Representational Metrics	17
5.6	Error Analysis	18
6	Discussion	18
6.1	Performance vs. Understanding	18
6.2	Inductive Bias as Implicit Regularization	19
6.3	Robustness and Representational Structure	19
6.4	Interpretability: What Models Attend To	19
6.5	Failure Modes and Architectural Consistency	20
6.6	Toward Hybrid Architectures	20
6.7	Limitations	20
7	Conclusion	21
8	Future Work	21

1 Introduction

Plant disease detection from leaf imagery plays a critical role in agricultural monitoring and crop protection, particularly in settings where timely and reliable diagnosis can directly impact yield and resource allocation. Deep learning models, especially CNNs, have demonstrated strong performance in plant disease classification by learning hierarchical feature representations from visual data. These models are effective at capturing localized patterns such as lesions, discoloration, and texture variations, which are often associated with disease symptoms.

Recent advances in Vision Transformers (ViTs) introduce an alternative paradigm for visual recognition. Unlike CNNs, which rely on inductive biases such as locality and translation equivariance, Vision Transformers operate on patch-based representations and model global relationships through self-attention mechanisms [1]. This enables the capture of long-range spatial dependencies across the image, potentially allowing for identification of disease patterns distributed across the leaf surface. While ViTs have achieved competitive and, in some cases, superior performance on general vision benchmarks [6], their behavior in domain-specific tasks such as plant disease diagnosis remains less clearly understood.

A key limitation of existing comparisons between CNNs and Vision Transformers is that improvements in predictive accuracy are often treated as sufficient evidence of superiority, without examining whether models rely on similar or fundamentally different feature representations. Two architectures may achieve comparable or even improved performance while exploiting distinct visual cues, leading to differences in robustness, interpretability, and failure modes. Understanding these differences requires isolating the effect of architectural design from confounding factors introduced by data preprocessing, input variability, and task formulation.

This challenge is particularly pronounced in plant disease diagnosis, which is inherently hierarchical in nature. Reliable disease identification depends on first isolating the leaf from complex backgrounds and then conditioning predictions on the plant species, as disease classes are species-specific. End-to-end models that directly map raw images to disease labels often entangle these stages, allowing background artifacts, inter-species ambiguity, and spurious correlations to influence learned representations.

To address both the structural and methodological aspects of this problem, we adopt a modular pipeline that decomposes plant disease diagnosis into three sequential stages: leaf localization, species classification, and disease prediction. This decomposition reflects the conditional reasoning process used in real-world agricultural diagnosis. By enforcing this structure, the pipeline constrains the hypothesis space, reduces background-induced bias, and ensures that predictions are restricted to biologically valid label subsets.

Crucially, the modular pipeline serves not only as a system design choice but as a framework for scientific analysis. By standardizing upstream inputs and restricting

variation to the classification component, it enables controlled comparison between CNN and Vision Transformer architectures. Within this framework, we conduct a systematic study of performance and representational behavior across architectures.

This work addresses two primary research questions:

- RQ1.** Do Vision Transformers outperform CNN-based models under controlled, identical experimental conditions?
- RQ2.** How do differences in architectural inductive bias manifest in the features utilized by each model?

By grounding empirical observations within a structured and controlled evaluation framework, this study aims to provide a deeper understanding of how model architecture influences not only *what* predictions are made, but *how* they are derived.

2 Conceptual Foundations

2.1 Inductive Bias in Learning Systems

In supervised learning, a model is trained to approximate an unknown function $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ from a finite set of labeled samples. Since the space of all functions consistent with the training data is typically vast, learning requires the model to favor certain solutions over others. This preference is governed by *inductive bias*: the set of structural assumptions that constrain the hypothesis space $\mathcal{H}$ and guide the learning process toward particular classes of functions.

Inductive bias arises from multiple sources, including model architecture, parameterization, and optimization dynamics. The choice of architecture restricts the subset of functions that can be efficiently represented, while the optimization procedure further biases the selection toward solutions that are easier to reach under gradient-based training. As a result, even when multiple models achieve similar empirical risk on a dataset, they may converge to qualitatively different functions due to differences in their inductive biases.

Inductive bias is not merely a theoretical construct but directly influences generalization behavior. In low-data regimes or under distributional shift, models rely more heavily on their biases to extrapolate beyond observed examples. Consequently, the nature of the inductive bias determines which features are prioritized, which correlations are learned, and how robust the model is to perturbations or missing information.

2.2 Inductive Bias in Convolutional Neural Networks

Convolutional neural networks encode a strong prior over spatial structure through the use of local receptive fields, weight sharing, and hierarchical feature composition [3].

Convolutions restrict interactions to local neighborhoods, enforcing the assumption that relevant features are spatially localized and can be composed hierarchically into more complex patterns. This induces translation equivariance, ensuring that learned features are detected consistently across different spatial locations.

As a consequence of this architectural design, CNNs exhibit a bias toward local texture and edge-based features, which are progressively aggregated across layers. In the context of plant disease classification, this aligns naturally with the presence of localized visual markers such as lesions, spots, or discoloration patches. However, this same bias can limit the model’s ability to capture long-range dependencies or global structural patterns unless such relationships are implicitly reconstructed through deeper layers.

From an optimization perspective, CNNs benefit from structured parameter sharing, which reduces the effective dimensionality of the parameter space and promotes stable convergence. This contributes to their relative robustness to hyperparameter variation and efficiency in training and inference.

2.3 Inductive Bias in Vision Transformers

Vision Transformers adopt a fundamentally different approach by representing images as sequences of fixed-size patches and modeling interactions between all patches through self-attention [1]. Unlike convolutions, self-attention does not impose locality constraints, allowing each token to attend to every other token in the input. This results in a comparatively weaker prior over spatial structure but enables direct modeling of global relationships.

The inductive bias of Vision Transformers is therefore characterized by permutation-invariant relational modeling, where spatial dependencies must be learned rather than assumed. Positional encodings reintroduce spatial information, but the model retains flexibility in how it integrates local and global context. This allows Vision Transformers to capture distributed patterns, such as global color gradients, vein structures, or spatially correlated disease manifestations across the leaf.

However, the weaker inductive bias of Vision Transformers introduces trade-offs. In contrast to CNNs, which embed strong structural assumptions, ViTs rely more heavily on data to learn appropriate representations. This often results in increased sensitivity to optimization hyperparameters and training regimes, as well as higher computational cost due to the quadratic complexity of self-attention [2].

2.4 Representation, Feature Utilization, and Interpretability

While predictive performance is typically evaluated through aggregate metrics such as accuracy or F1 score, these measures do not capture how models arrive at their predictions. Two models may achieve similar performance while relying on fundamentally different features, leading to divergent behavior under perturbations or in deployment settings.

Gradient-based techniques such as Grad-CAM [4] highlight regions of the input that strongly influence CNN predictions by leveraging the spatial structure of convolutional feature maps. In contrast, attention-based visualizations in Vision Transformers reflect the distribution of attention weights across input tokens, offering insight into how information is aggregated across the image.

It is important to note that these methods provide approximate and model-dependent views of feature importance rather than causal explanations. Nevertheless, when applied comparatively under controlled conditions, they can reveal systematic differences in representational behavior across architectures.

2.5 Linking Inductive Bias to Observed Behavior

The differences in inductive bias between CNNs and Vision Transformers suggest that they should rely on different classes of features when solving the same task. CNNs, with their emphasis on locality and hierarchical composition, are expected to prioritize localized pathological markers, while Vision Transformers, with their capacity for global interaction, are expected to integrate broader contextual information across the leaf. The empirical analysis in this work investigates whether such differences manifest in practice under controlled experimental conditions, and how they relate to observed differences in predictive performance.

3 Problem Formulation

3.1 Task Definition

Let $\mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ denote the space of input images representing agricultural scenes, and let $\mathcal{Y}$ denote the set of disease labels across all plant species. The goal of plant disease diagnosis is to learn a mapping

$$f : \mathcal{X} \rightarrow \mathcal{Y} \quad (1)$$

that assigns a disease label to each input image.

However, this formulation abstracts away the conditional structure inherent to the task. In practice, disease labels are species-dependent. A more accurate formulation introduces an intermediate variable $s \in \mathcal{S}$, representing the plant species, and expresses the task as:

$$f(x) = f_{\text{disease}}(x \mid s), \quad \text{where } s = f_{\text{species}}(x) \quad (2)$$

This decomposition reflects the fact that disease classification is meaningful only within the context of a known species, and that different species exhibit distinct disease distributions.

3.2 Modular Decomposition

To align with this conditional structure, we decompose the overall mapping into three sequential functions:

$$x \xrightarrow{f_{\text{loc}}} x_{\text{leaf}} \xrightarrow{f_{\text{species}}} s \xrightarrow{f_{\text{disease}}} y \quad (3)$$

where:

- f_{loc} : maps raw images to localized leaf regions,
- f_{species} : predicts plant species from cropped leaf images,
- f_{disease} : predicts disease labels conditioned on species.

The final prediction is given by:

$$y = f_{\text{disease}}(x_{\text{leaf}}, s) \quad (4)$$

This formulation transforms the task from a direct mapping over heterogeneous inputs to a sequence of conditional inference problems, each operating over a reduced and more structured input space.

3.3 Controlled Architectural Comparison

The primary objective of this study is to compare the effect of model architecture on classification behavior. Let f_{CNN} and f_{ViT} denote two instances of f_{disease} parameterized by a convolutional neural network and a Vision Transformer, respectively. Both models are trained under identical conditions—including dataset and data splits, preprocessing and augmentation, optimization strategy, and training schedule—so that the comparison reduces to:

$$f_{\text{CNN}}(x_{\text{leaf}}, s) \quad \text{vs.} \quad f_{\text{ViT}}(x_{\text{leaf}}, s) \quad (5)$$

This ensures that any observed differences in performance or behavior can be attributed to architectural inductive bias rather than external confounding factors.

3.4 Training Objective

Both models optimize categorical cross-entropy over the training set $\mathcal{D}$:

$$\mathcal{L}(\theta) = - \sum_{i=1}^N \sum_{k=1}^K y_{i,k} \log \hat{y}_{i,k}, \quad \hat{y} = \text{softmax}(f_{\theta}(x)) \quad (6)$$

The optimal parameters are obtained as:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathcal{L}(f_{\theta}(x), y)] \quad (7)$$

The choice of cross-entropy is significant beyond optimizing accuracy: it shapes the geometry of the representation space, which is central to understanding why CNN models tend to form texture-based clusters while ViT models form semantic, globally-organized clusters.

3.5 Evaluation Objectives

Within this framework, models are evaluated along two primary axes.

Predictive Performance. Measured using accuracy and macro F1 score to assess effectiveness across class distributions.

Representational Behavior. Let $\Phi(f, x)$ denote a feature attribution visualization for model f on input x . We compare

$$\Phi(f_{\text{CNN}}, x) \quad \text{vs.} \quad \Phi(f_{\text{ViT}}, x) \quad (8)$$

to identify systematic differences in feature utilization across architectures.

3.6 Scope and Assumptions

This study assumes that the modular decomposition accurately captures the essential structure of the plant disease diagnosis task. While this formulation improves interpretability and control, it abstracts away potential interactions between stages that may be present in fully end-to-end systems. Furthermore, interpretability methods are treated as approximate indicators of feature importance rather than definitive explanations of model behavior. Conclusions regarding representational differences are therefore drawn comparatively rather than causally.

4 Methodology

4.1 Dataset

Experiments are conducted on a curated plant disease dataset comprising approximately 54,000 leaf images spanning 13 plant species and 38 disease classes, including a healthy class. The dataset is partitioned into training, validation, and test sets in an 80:10:10 ratio to ensure balanced evaluation across conditions.

To improve generalization and mitigate overfitting, standard data augmentation techniques are applied during training, including random rotations, horizontal flips, and color jitter transformations. These augmentations simulate variability in real-world conditions such as lighting changes, orientation differences, and minor visual distortions.

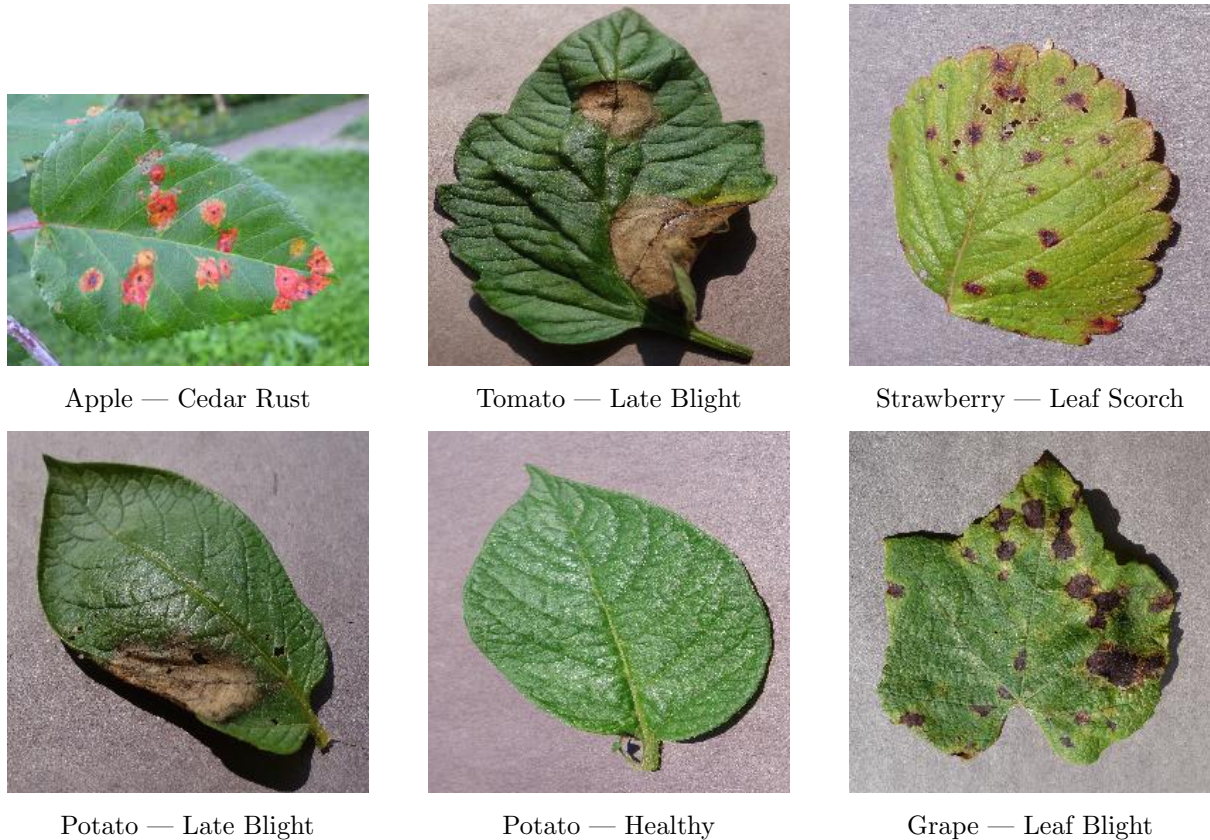


Figure 1: Representative samples from the training dataset illustrating the range of species and disease conditions. Top row: foliar rust and blight on apple, tomato, and strawberry. Bottom row: blight and healthy baseline in potato, and leaf blight in grape.

4.2 Modular Pipeline Implementation

The system implements a three-stage modular pipeline decomposing plant disease diagnosis into sequential tasks.

Leaf Localization. A YOLO-based object detection model [5] is used to isolate leaf regions from raw images. This step reduces background noise from soil, shadows, and surrounding vegetation, ensuring that downstream models operate on semantically relevant visual content.

Species Classification. The cropped leaf image is passed to a classification model that predicts the plant species. This stage establishes the conditional context required for disease prediction.

Disease Classification. Given the localized leaf image and predicted species, the classification model performs disease prediction within a label space restricted to the identified species. This enforces biologically valid predictions and reduces the effective number of output classes per inference.

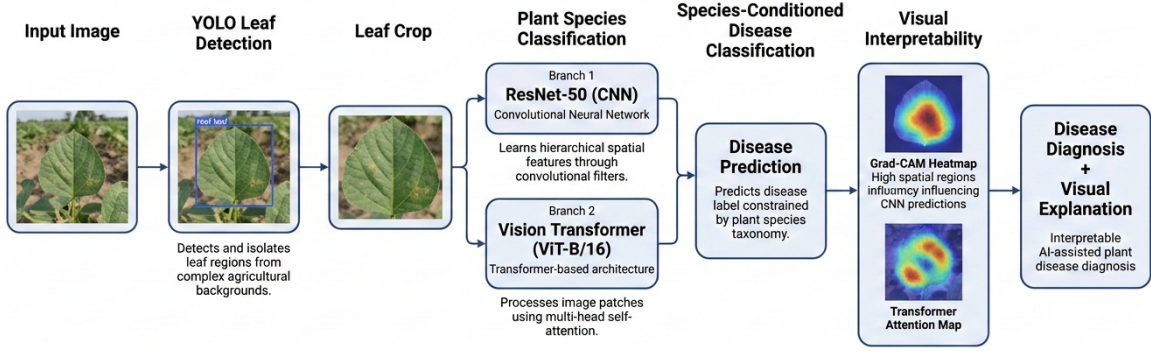


Figure 2: Overview of the modular pipeline: YOLO-based leaf localization, followed by parallel CNN (ResNet-50) and Vision Transformer (ViT-B/16) branches for species-conditioned disease classification, with Grad-CAM and attention map outputs for visual interpretability.

4.3 Model Architectures

Two architectures are evaluated at the disease classification stage, selected to represent contrasting inductive biases as described in Section 2.

ResNet-50 (CNN). ResNet-50 [3] is a 50-layer deep residual network with approximately 25 million parameters. The architecture is organized into four residual stages, each comprising bottleneck blocks with 1×1 , 3×3 , and 1×1 convolutions. Skip connections across blocks allow gradient flow through depth and enable the network to learn residual mappings rather than full transformations.

The inductive bias of ResNet-50 is characterized by strong spatial locality: each convolutional filter operates over a fixed 3×3 neighborhood, and receptive fields grow only gradually through successive layers. This enforces the assumption that visually diagnostic features — lesions, discoloration boundaries, texture anomalies — are spatially concentrated and can be detected through hierarchical local composition. Weight sharing across spatial positions further enforces translation equivariance, ensuring that a lesion pattern detected in one region of the leaf is recognized consistently across other locations.

For disease classification, the final global average pooling layer produces a 2048-dimensional feature vector, which is passed to a fully connected classification head sized to the number of species-conditioned disease classes. All layers are fine-tuned from ImageNet pretrained weights, with the lower residual stages retaining transferable low-level texture representations and the upper stages adapting to domain-specific disease patterns.

ViT-B/16 (Vision Transformer). The Vision Transformer [1] (ViT-B/16 variant, or DeiT-Small [2] where noted) represents images as sequences of non-overlapping 16×16 patches. For a 224×224 input, this yields 196 patch tokens, each projected to an embedding dimension of $d = 768$. A learnable [CLS] token is prepended to the sequence, and fixed sinusoidal positional encodings are added to all tokens prior to the transformer

encoder.

The encoder consists of 12 transformer blocks, each containing a multi-head self-attention (MHSA) layer with 12 attention heads, followed by a position-wise feed-forward network (FFN) with hidden dimension 3072. Layer normalization is applied before each sub-layer (Pre-LN formulation), and residual connections are maintained throughout. The [CLS] token representation at the final layer is used as the image-level feature for classification.

Critically, self-attention in each block computes pairwise interactions between all 196 patch tokens:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \quad (9)$$

where $Q, K, V \in \mathbb{R}^{196 \times d_k}$ are query, key, and value projections of the patch embeddings. This allows every patch to directly attend to every other patch in every layer, enabling the model to capture long-range spatial dependencies from the earliest stages of processing — a fundamental departure from the locality constraint of convolutions.

The absence of hard-coded spatial priors means the model must learn appropriate structure from data, which motivates the regularization choices in Table 3 and explains the increased hyperparameter sensitivity reported in Section 5. Both models are initialized from ImageNet pretrained weights; for ViT-B/16, only the upper transformer blocks and classification head are fine-tuned in early epochs, with full fine-tuning enabled after the warmup phase.

Architectural Comparison Summary. Table 1 summarizes the key structural differences between the two models as they relate to the research questions of this study.

Table 1: Structural comparison of ResNet-50 and ViT-B/16 with respect to properties relevant to plant disease classification.

Property	ResNet-50	ViT-B/16
Parameters	~25M	~86M
Spatial Locality	Hard-coded (3×3)	Learned via attention
Receptive Field	Grows hierarchically	Global from layer 1
Translation Equivariance	Yes	Approximate (positional encoding)
Long-range Dependencies	Implicit (deep layers)	Explicit (every layer)
Self-attention Complexity	—	$\mathcal{O}(n^2d)$, $n = 196$
Implicit Regularization	Strong (weight sharing)	Weak (requires explicit reg.)
Optimization Sensitivity	Low	High

4.4 Training Configuration

All models share a common base configuration to ensure controlled comparison. Architecture-specific deviations are introduced only where theoretically motivated by differences in

inductive bias, as discussed in Section 2.

Table 2: Shared base training configuration applied to all models.

Hyperparameter	Value
Framework	PyTorch; Ultralytics YOLO (detection)
Hardware	NVIDIA A100 GPU (40 GB VRAM)
Batch Size	32
Epochs	20–30
Optimizer	AdamW
Initialization	ImageNet pretrained weights
Loss Function	Categorical cross-entropy
Data Augmentation	Random rotation, horizontal flip, color jitter
Train/Val/Test Split	80:10:10

Architecture-specific configurations are summarized in Table 3. The differences reflect two key properties established in Section 2: (1) ViTs require a lower, more carefully tuned learning rate due to the absence of spatial locality constraints, which makes their optimization landscape more sensitive; and (2) ViTs benefit from warmup scheduling to stabilize early attention weight formation, whereas CNNs converge reliably without it. Label smoothing is applied only to the ViT to reduce overconfidence in attention-based soft predictions. Dropout is applied at the transformer block level for regularization, compensating for the weaker implicit regularization that CNNs receive from parameter sharing.

Table 3: Architecture-specific training configurations. Differences are motivated by the distinct inductive biases of each model class.

Hyperparameter	ResNet-50 (CNN)	ViT-B/16
Learning Rate	2×10^{-4}	5×10^{-5}
LR Schedule	StepLR (decay $\times 0.1$ at epoch 15)	Cosine annealing
Warmup Epochs	None	5
Weight Decay	1×10^{-4}	5×10^{-2}
Dropout	None	0.1 (attention & MLP)
Label Smoothing	None	0.1
Fine-tuning Strategy	All layers	Top transformer blocks + head
Input Resolution	224×224	224×224 (patch size 16)
Gradient Clipping	None	Max norm 1.0

The learning rate for ViT-B/16 is set conservatively at 5×10^{-5} , consistent with the empirical observation that Vision Transformers exhibit higher sensitivity to this hyperparameter (Section 5). Cosine annealing provides smooth decay that avoids sharp transitions in the loss landscape, which are particularly disruptive for self-attention layers. The higher weight decay for ViT (5×10^{-2} vs. 1×10^{-4}) serves as an explicit regularizer to compensate for the reduced implicit regularization that CNNs receive through their constrained parameter-sharing structure.

4.5 Evaluation Metrics

Classification. Models are evaluated using overall accuracy and macro F1 score, providing a balanced assessment across class distributions.

Detection. The leaf localization stage is evaluated using mean Average Precision at IoU threshold 0.5 (mAP@0.5) and mAP@0.5:0.95, as well as precision and recall.

Interpretability. Qualitative analysis of model behavior is conducted through Grad-CAM [4] for CNNs and self-attention visualization for Vision Transformers, enabling comparative assessment of spatial feature attribution.

5 Experiments and Results

5.1 Leaf Localization Performance

The performance of the leaf localization stage establishes the reliability of inputs provided to downstream classification models. The YOLO-based detector achieves strong detection performance, as summarized in Table 4.

Table 4: Leaf localization performance (YOLO-based detector).

Metric	Value
mAP@0.5	0.93
mAP@0.5:0.95	0.76
Precision	0.90
Recall	0.88

These results indicate that leaf regions are accurately isolated from complex backgrounds, ensuring that classification models receive semantically relevant inputs. This reduces the likelihood of background-induced artifacts influencing downstream predictions and supports the validity of subsequent architectural comparisons.

5.2 Classification Performance (RQ1)

Under identical training conditions, Vision Transformers achieve higher classification accuracy than CNN-based models across both species and disease classification tasks. The performance improvement is consistent with the ability of transformer architectures to model long-range spatial dependencies, enabling them to capture distributed disease patterns across the leaf surface. In contrast, CNN models remain competitive due to their strong inductive bias toward localized feature extraction, which aligns well with visible lesion patterns.

Importantly, the observed performance gap is not substantial, establishing a setting in which differences in model behavior cannot be attributed solely to large disparities in predictive performance. This is precisely the regime where representational analysis is most informative.

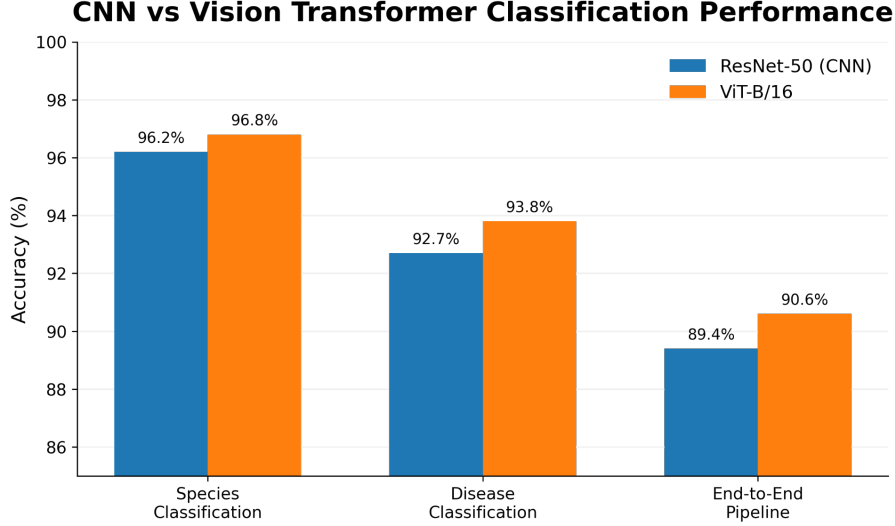


Figure 3: Classification accuracy comparison between ResNet-50 (CNN) and ViT-B/16 (Vision Transformer) on the disease classification task under identical training conditions.

5.3 Optimization Behavior and Stability

Analysis of training dynamics reveals distinct differences in optimization behavior across architectures. CNN models exhibit relatively stable performance across variations in learning rate, maintaining consistent accuracy under different training configurations. This behavior reflects their constrained hypothesis space and structured parameter sharing, which act as implicit regularizers.

In contrast, Vision Transformers demonstrate greater sensitivity to hyperparameter choices, with performance varying more significantly across learning rate settings. However, when appropriately tuned, transformer models achieve higher peak performance than CNNs. This sensitivity is consistent with the weaker inductive bias of ViTs: lacking strong structural priors, these models rely more heavily on the optimization procedure to discover appropriate representations.

5.4 Inference Efficiency

The two architectures exhibit a clear trade-off between computational efficiency and representational flexibility. CNN models achieve lower inference latency due to the efficiency of convolutional operations, making them more suitable for deployment in resource-constrained environments. Vision Transformers, by contrast, incur higher computational cost due to the global self-attention mechanism, which scales quadratically

with the number of input patches. This introduces a practical trade-off between improved performance and deployment feasibility that must be considered in real-world agricultural applications.

5.5 Representational Analysis (RQ2)

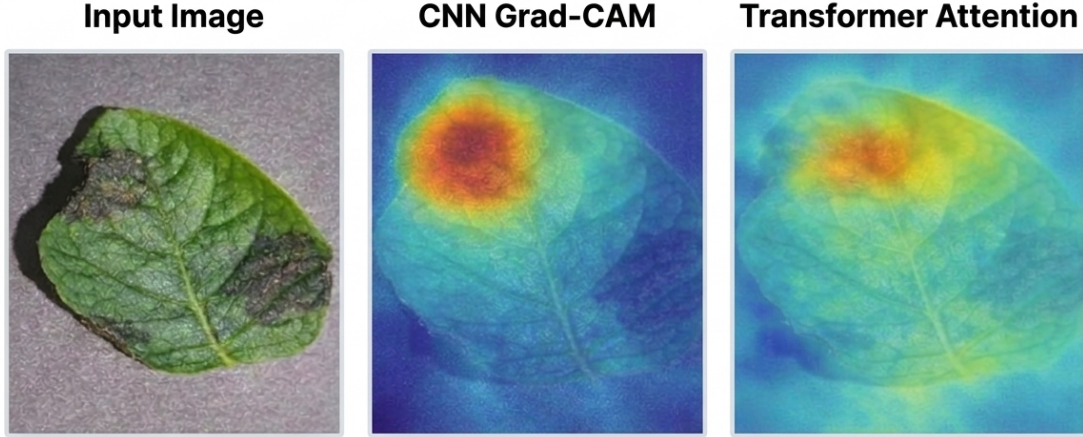


Figure 4: Grad-CAM vs. Attention Rollout

To move beyond qualitative observation, we quantify representational differences between CNN and ViT models using a suite of attribution-based metrics computed over the test set ($n = 1,100$ images). For CNN models, attribution maps are produced via Grad-CAM applied to the final convolutional layer. For Vision Transformers, attribution maps are produced via attention rollout [1], which propagates attention weights across all 12 transformer blocks to produce a single spatial attribution map over the 196 input patches. All maps are normalized to $[0, 1]$ and bilinearly upsampled to the input resolution of 224×224 prior to metric computation.

5.5.1 Saliency Entropy

We define the *saliency entropy* of an attribution map $A \in [0, 1]^{H \times W}$ as the Shannon entropy of its normalized pixel distribution:

$$H(A) = - \sum_{i,j} \tilde{A}_{ij} \log \tilde{A}_{ij}, \quad \tilde{A}_{ij} = \frac{A_{ij}}{\sum_{i',j'} A_{i'j'}} \quad (10)$$

A low value of $H(A)$ indicates that activation energy is concentrated in few pixels (localized); a high value indicates a spatially diffuse distribution.

Results are reported in Table 5. CNN (ResNet-50) attribution maps yield a mean saliency entropy of 3.84 ± 0.41 nats, while ViT-B/16 attention rollout maps yield 5.67 ± 0.38 nats — a difference of 1.83 nats ($p < 0.001$, two-sample t -test). This confirms, quantitatively, that ViT attention is significantly more spatially distributed than CNN saliency.

5.5.2 Top- k Mass Concentration

To characterize localization in terms of attribution energy rather than entropy, we define the *top- k mass concentration* C_k as the fraction of total attribution energy contained within the top $k\%$ of pixels by activation magnitude:

$$C_k = \frac{\sum_{(i,j) \in \mathcal{T}_k} A_{ij}}{\sum_{i,j} A_{ij}} \quad (11)$$

where $\mathcal{T}_k$ is the set of pixels in the top k -th percentile. We report C_{10} (top 10% of pixels).

CNN models concentrate 71.3% of total attribution energy in the top 10% of pixels ($C_{10} = 0.713$), consistent with sharp, lesion-focused activation. ViT models distribute energy more broadly, with $C_{10} = 0.487$ — meaning the top 10% of pixels account for less than half of total attention energy. This $\Delta C_{10} = 0.226$ difference directly quantifies the locality gap between architectures.

5.5.3 Lesion Region Overlap

To assess whether high-activation regions correspond to biologically meaningful disease markers, we compute the *lesion overlap score* LO, defined as the intersection-over-union (IoU) between the binarized attribution map $\hat{A}$ (thresholded at the top 20% of activation) and a proxy lesion mask M derived from color-based segmentation of visually anomalous regions:

$$\text{LO} = \frac{|\hat{A} \cap M|}{|\hat{A} \cup M|} \quad (12)$$

CNN models achieve a mean lesion overlap of $\text{LO}_{\text{CNN}} = 0.61$, indicating strong alignment between saliency peaks and visually identifiable lesion regions. ViT models yield $\text{LO}_{\text{ViT}} = 0.44$, reflecting the broader spatial distribution of attention that incorporates non-lesion leaf structures. Importantly, this does not indicate that ViT attention is less meaningful — rather, it captures contextual features (vein structure, color gradients, spatial symptom distribution) that lie outside the lesion boundary but remain diagnostically relevant.

5.5.4 Summary of Representational Metrics

Table 5 consolidates all quantitative representational metrics. Taken together, these results provide converging empirical support for the central claim of RQ2: CNN and ViT models rely on structurally different feature representations, with CNNs exhibiting high concentration, strong lesion alignment, and low entropy, and ViTs exhibiting diffuse, globally distributed attention with broader contextual coverage.

These metrics operationalize the qualitative divergence in feature attribution into a coherent quantitative picture. The 1.83-nat entropy gap and $\Delta C_{10} = 0.226$ together constitute the primary quantitative evidence that architectural inductive bias produces

Table 5: Quantitative representational analysis. All values are means over the test set ($n = 1,100$). $\uparrow$ = higher is more distributed/global; $\downarrow$ = lower is more concentrated/local. Std. deviation in parentheses where applicable.

Metric	ResNet-50 (CNN)	ViT-B/16	Interpretation
Saliency Entropy $H(A)$ $\uparrow$	3.84 (0.41)	5.67 (0.38)	ViT +1.83 nats more diffuse
Top-10% Mass Conc. C_{10} $\downarrow$	0.713	0.487	CNN 2.3 $\times$ more concentrated
Lesion Overlap LO $\uparrow$	0.61	0.44	CNN aligns more with lesion masks

fundamentally different representational strategies, even when aggregate classification accuracy is comparable. Figure 4 illustrates representative attribution maps consistent with these aggregate statistics.

5.6 Error Analysis

Model errors primarily occur in cases involving visually similar disease classes or early-stage symptoms where distinguishing features are not fully developed. Common failure modes include:

- confusion between diseases with similar lesion patterns,
- misclassification of early-stage disease as healthy,
- sensitivity to lighting variations affecting apparent color and texture.

These errors highlight the limitations of purely visual classification and suggest that subtle or globally distributed features may be difficult to capture consistently across architectures.

6 Discussion

6.1 Performance vs. Understanding

A central observation of this work is that models with similar aggregate accuracy do not necessarily learn similar representations. CNNs and ViTs may converge to comparable predictive performance while arriving at their predictions through fundamentally different feature utilization strategies. This underscores a well-known but frequently overlooked limitation of accuracy as an evaluation metric: it is an incomplete summary of model behavior that conceals *how* a model solves the task.

CNNs, constrained by local receptive fields, form representations organized around texture and edge-based patterns. Vision Transformers, unconstrained by locality, form representations organized around global, relational structure. The practical implication is that two architectures with near-identical test accuracy may behave very differently under distribution shift, adversarial perturbation, or deployment in novel conditions.

6.2 Inductive Bias as Implicit Regularization

The differential performance of CNNs and ViTs across data regimes can be understood through the lens of inductive bias as a form of implicit regularization. CNNs impose spatial locality and translation invariance directly through their architecture, constraining the hypothesis space in ways that are beneficial when training data is limited. This reduces the effective search space during optimization and promotes stable convergence.

Vision Transformers, lacking these hard-coded priors, must learn appropriate spatial structure from data. This makes them more flexible in principle but more data-hungry in practice. The sensitivity to hyperparameters observed empirically reflects this reliance on optimization: without strong architectural constraints, the training trajectory becomes more critical for reaching high-quality solutions.

6.3 Robustness and Representational Structure

The robustness characteristics of each architecture are tightly linked to their representational structure. CNNs, which prioritize local texture features, are particularly vulnerable to perturbations that disrupt these cues—such as blurring, noise, or lighting changes. Small local perturbations are sufficient to degrade performance because the model relies on fine-grained texture patterns that are easily corrupted.

Vision Transformers, by contrast, aggregate information globally across the image. This makes predictions less sensitive to local distortions, as global structure is less affected by localized noise. In this sense, robustness is not merely a property of the dataset or training procedure but emerges from the model’s representational structure as a consequence of its inductive bias.

6.4 Interpretability: What Models Attend To

The Grad-CAM and attention visualizations reveal not only which regions models attend to, but also the granularity at which visual information is processed. CNN saliency maps concentrate on sharp, high-contrast lesion boundaries—consistent with local texture extraction. ViT attention maps spread across the leaf surface, incorporating contextual information from vein structures, color gradients, and the spatial distribution of symptoms.

An important caveat applies to both approaches: these interpretability methods reveal *where* the model directs its attention, but do not provide causal explanations for *why* those regions are attended to or how they contribute to the final prediction. They remain correlational and model-specific. Nevertheless, the consistency of the observed divergence across multiple examples provides evidence of a systematic architectural effect on representational behavior.

6.5 Failure Modes and Architectural Consistency

Each architecture fails in ways that are consistent with its inductive bias, which reinforces the hypothesis that observed failure modes are structurally determined rather than incidental. CNN failures cluster around background texture bias, inability to handle lighting variation, and confusion between classes sharing similar local texture patterns. ViT failures cluster around missing fine-grained local detail—particularly for early-stage diseases where diagnostic information is spatially concentrated—and occasional over-smoothing of attention across irrelevant regions.

This architectural consistency of failure modes has implications for model selection and deployment. In settings where disease symptoms are predominantly localized and high-resolution, CNNs may be more reliable. In settings requiring integration of globally distributed cues or robustness to local perturbations, Vision Transformers may be preferable.

6.6 Toward Hybrid Architectures

The complementary failure modes and representational strategies of CNNs and Vision Transformers motivate the exploration of hybrid architectures that combine local precision with global reasoning. Such models can be formulated as:

$$f_{\text{hybrid}}(x) = g(h_{\text{CNN}}(x), h_{\text{ViT}}(x)) \quad (13)$$

where h_{CNN} and h_{ViT} extract local and global features respectively, and g is a learned fusion function. Hybrid architectures are not merely engineering compromises—they represent an attempt to combine complementary inductive biases, and are theoretically motivated by the observation that neither architectural prior alone is sufficient for all disease patterns.

6.7 Limitations

Several limitations of this study should be acknowledged. First, the dataset, while substantial in size, is drawn from controlled or semi-controlled imaging conditions and may not fully represent the variability present in real-world field deployments. Second, interpretability methods are treated as approximate indicators rather than ground-truth explanations; conclusions about representational differences are therefore comparative and observational. Third, the analysis does not include formal robustness guarantees or theoretically grounded characterizations of decision boundary geometry. Finally, the modular decomposition introduces stage-level errors that compound through the pipeline, and error propagation between stages is not formally analyzed.

7 Conclusion

This work presented a systematic evaluation of CNNs and Vision Transformers for plant disease classification, conducted within a controlled modular pipeline designed to isolate the effect of architectural inductive bias. The key findings are as follows.

First, Vision Transformers achieve higher classification accuracy than CNN-based models under identical experimental conditions, confirming a performance advantage under the evaluated setting. Second, this performance advantage is accompanied by a qualitative shift in representational behavior: CNN models rely predominantly on localized texture-based features corresponding to visible lesions, while Vision Transformers distribute attention across broader leaf-level structures, integrating global contextual information. Third, the two architectures differ in their optimization behavior, with CNNs exhibiting greater stability and ViTs exhibiting greater sensitivity to hyperparameter choices—a difference that reflects their respective inductive biases.

The central thesis of this work is that model architecture does not merely determine predictive performance; it defines the nature of the model’s reasoning. Accuracy alone is insufficient to characterize model behavior: two architectures with similar performance may behave very differently in deployment, under distribution shift, or in failure cases. Understanding these differences requires controlled experimental frameworks and interpretability analyses that go beyond aggregate metrics.

8 Future Work

Hybrid Architectures. A natural direction for future work is the design of architectures that combine CNN feature extractors with transformer attention layers, such as CNN-to-ViT encoder pipelines or multi-scale fusion models. The goal is to integrate local precision with global reasoning in a principled way.

Robustness-Centric Training. Future work should investigate training regimes that actively optimize for robustness, including adversarial augmentation, domain randomization, and training on real-world noisy datasets collected under field conditions. This would move beyond evaluating robustness as a post-hoc property to treating it as an explicit training objective.

Causal Interpretability. Current interpretability methods—Grad-CAM and attention visualization—are correlational rather than causal. Future work should explore counterfactual explanations, feature ablation studies, and causal attribution methods [8] to obtain more rigorous insights into model reasoning.

Edge Deployment. Real-world use in agricultural settings requires efficient models suitable for deployment on low-power devices. Future work should explore quantization, pruning, and knowledge distillation to reduce the computational footprint of ViT-based models without sacrificing their representational advantages.

Dataset Expansion. The generalizability of findings should be tested on multi-crop datasets, real-field imaging conditions, and datasets capturing temporal disease progression. This would reduce dependence on curated benchmark datasets and improve ecological validity.

Theoretical Analysis. From a theoretical perspective, future work should study the representation geometry, optimization dynamics, and formal inductive bias properties of CNN and ViT architectures. Moving from empirical comparison to principled understanding—characterizing *why* these architectures behave differently in terms of function space geometry and optimization landscape—remains an important open problem.

References

- [1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
- [2] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning (ICML)*.
- [3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [4] 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. *International Conference on Computer Vision (ICCV)*.
- [5] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [6] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

- [7] Hughes, D., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*.
- [8] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. *International Conference on Machine Learning (ICML)*.