Advanced Out-of-Distribution Detection Frameworks for Fine-Grained Plant Disease Diagnosis: A Synthesis of Vision Transformers, Foundation Models, and Parameter-Efficient Adaptation

Executive Summary

The deployment of automated diagnostic systems in precision agriculture is contingent upon their reliability in open-world environments. While deep learning has achieved superhuman performance in closed-set classification of plant diseases, the operational reality of agricultural fields introduces a vast array of Out-of-Distribution (OOD) stimuli—ranging from novel pathogen strains and abiotic stressors to non-plant artifacts and sensor anomalies. The inability of standard classifiers to robustly reject these anomalies constitutes a critical safety vulnerability; a high-confidence misclassification of a novel, virulent blight as a benign nutrient deficiency can lead to catastrophic crop loss.

This report presents a comprehensive technical analysis of the state-of-the-art methodologies for OOD detection, specifically tailored to the architectural paradigm of Vision Transformers (ViTs) and the domain constraints of Fine-Grained Visual Categorization (FGVC). We critically examine the transition from classical Extreme Value Theory (EVT) approaches, such as OpenMax, to modern geometric methods operating in the latent spaces of Foundation Models like DINOv2. A central focus is placed on the interaction between Parameter-Efficient Fine-Tuning (PEFT), specifically Low-Rank Adaptation (LoRA), and uncertainty estimation. Our analysis synthesizes empirical evidence to demonstrate that unmerged LoRA embeddings, when analyzed via Mahalanobis Distance, offer a superior mechanism for detecting "near-OOD" samples—those semantically proximate anomalies that plague fine-grained disease classification. Furthermore, we explore the interpretability of these decisions through multi-scale attention mechanisms and Class Activation Mapping (CAM), proposing a unified framework that balances sensitivity, robustness, and explainability.

---

1. The Operational Imperative: Reliability in Fine-Grained Agricultural AI

1.1 The Fine-Grained Visual Categorization (FGVC) Challenge in

Pathology

Plant disease diagnosis is a quintessential Fine-Grained Visual Categorization (FGVC) problem. Unlike generic object recognition (e.g., distinguishing a cat from a dog), FGVC requires discriminating between subclasses that share a high degree of structural similarity. In the context of phytopathology, a "healthy" tomato leaf, a leaf with "Early Blight" (*Alternaria solani*), and a leaf with "Septoria Leaf Spot" (*Septoria lycopersici*) share the same global geometry, color palette, and biological morphology. The discriminative features are often minute, localized textural anomalies—concentric rings in a necrotic spot versus water-soaked lesions.[1]

This high intra-class variance (due to varying growth stages, lighting, and leaf angles) and low inter-class variance creates a perilous landscape for OOD detection. A standard Convolutional Neural Network (CNN) or Vision Transformer (ViT) trained on a closed set of 10 diseases will essentially partition the high-dimensional feature space into 10 regions. When presented with an 11th, unknown disease (OOD), the model forces the input into one of the existing partitions. Because the global structure (a leaf) is In-Distribution (ID), the model often assigns high confidence to the prediction based on shared background features, a phenomenon known as the "overconfidence" problem.

1.2 The Shift from CNNs to Vision Transformers

Historically, CNNs dominated this field, leveraging their inductive bias for local texture to identify lesions. However, the field is rapidly shifting toward Vision Transformers (ViTs). ViTs, which process images as sequences of patch tokens, utilize self-attention mechanisms capable of modeling long-range dependencies. This is theoretically advantageous for plant disease detection, where the spatial distribution of lesions (e.g., scattered spots vs. marginal necrosis) is diagnostic.[3]

However, ViTs introduce new complexities for OOD detection.

- **Feature Uniformity:** The Layer Normalization (LayerNorm) inherent in ViTs tends to produce feature representations that are more uniform in magnitude and distribution compared to the unnormalized activations of CNNs.
- **Shape Bias:** ViTs exhibit a stronger "shape bias" compared to the "texture bias" of CNNs. While this improves robustness to occlusion, it can make the model less sensitive to the textural anomalies that distinguish different pathologies, potentially clustering distinct diseases closer together in the latent space.[5]
- **OOD Behavior:** Empirical studies suggest that standard post-hoc detection methods optimized for CNNs (like ODIN or React) often degrade in performance when applied directly to ViTs without modification.[7]

1.3 The Scope of Analysis

This report dissects specific mechanisms to address these challenges:

1. **Geometric Methods:** The Mahalanobis Distance and its variants (RMD, Mahalanobis++), which leverage the covariance structure of the feature space.
2. **Probabilistic Methods:** OpenMax and Weibull-based tail modeling.
3. **Foundation Model Strategies:** The specific utility of DINOv2 and the trade-off between linear probing and non-parametric nearest-neighbor evaluations.
4. **Adaptation-Based Detection:** The novel exploitation of LoRA parameters as uncertainty sensors.
5. **Validation:** The use of attention mechanisms and CAM to visually verify OOD decisions.

---

2. Geometric Approaches: Mahalanobis Distance and Its Evolutions

The geometric interpretation of neural feature spaces provides the most mathematically grounded framework for OOD detection. The core hypothesis is that a well-trained network maps ID data to a union of compact, low-dimensional manifolds (approximated as Gaussians), and OOD data falls into the low-density regions between or outside these manifolds.

2.1 Theoretical Foundations of Mahalanobis Distance (MD)

The Mahalanobis Distance (MD) is a generalized distance metric that accounts for the correlations between variables in a dataset. Unlike the Euclidean distance, which assumes that features are uncorrelated and have unit variance (isotropic), MD "whitens" the data based on the empirical covariance matrix.[8]

Formally, consider a pre-trained feature extractor $f(x)$ that maps an input image x to a feature vector $z \in \mathbb{R}^d$ in the penultimate layer. We model the distribution of features for each class $c \in \{1, \dots, C\}$ as a multivariate Gaussian distribution $\mathcal{N}(\mu_c, \Sigma)$.

The class-conditional mean is estimated as:

\hat{\mu}_c = \frac{1}{N_c} \sum_{i: y_i = c} f(x_i)$$To ensure robust estimation, especially in high-dimensional spaces where the number of samples per class (N_c) might be small relative to the dimensionality (d), we typically assume a **tied covariance matrix** Σ shared across all classes. This is calculated by pooling the sample covariances:$$\hat{\Sigma} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i: y_i = c} (f(x_i) - \hat{\mu}_c)(f(x_i) - \hat{\mu}_c)^T

The OOD score for a test sample x is defined as the minimum squared distance to any class centroid, scaled by the inverse covariance:

$$M(x) = \min_c (f(x) - \hat{\mu}_c)^T \hat{\Sigma}^{-1} (f(x) - \hat{\mu}_c)$$

2.1.1 Why MD Outperforms Softmax

The Softmax function in the final layer is a projection that compresses the high-dimensional feature vector into a probability simplex. This compression inevitably results in information loss. The Softmax logits effectively represent the distance to the decision boundary (hyperplane), not the distance to the class prototype. A sample can be extremely far from the centroid (an outlier) but still be far from the decision boundary (high confidence), leading to the "high-confidence fool" problem. MD, by operating in the feature space, captures the distance to the density mode, providing a direct measure of "typicality".[8]

2.2 The "Simple Fix": Relative Mahalanobis Distance (RMD)

Despite the theoretical elegance of MD, it suffers from a critical failure mode in "Near-OOD" scenarios—situations where the OOD samples share significant semantic overlap with the ID data. This is precisely the case in plant disease detection, where a "Tomato Yellow Leaf Curl" image (OOD) looks structurally identical to a "Tomato Mosaic Virus" image (ID) except for specific coloration patterns.[11]

2.2.1 The Background Confounding Problem

In deep neural networks, the magnitude (norm) of the feature vector carries significant information about the "background" content of the image. For instance, the presence of a leaf shape and green pixels triggers a baseline level of activation across the network filters. This "background signal" contributes to the total distance $M(x)$. When both ID and Near-OOD samples share this background, their $M(x)$ scores become indistinguishable, as the shared background distance dominates the subtle class-specific distance.[12]

2.2.2 The RMD Methodology

Ren et al. [11] proposed the Relative Mahalanobis Distance (RMD) to isolate the class-specific signal. The method assumes that the feature vector is composed of a class-specific component and a class-agnostic (background) component. To estimate the background contribution, RMD fits a second Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$ to the **entire training set**, ignoring class labels. The RMD score is the difference between the class-specific distance and the background distance:

$$RMD(x) = M(x) - M_0(x)$$

where $M_0(x) = (f(x) - \hat{\mu}_0)^T \hat{\Sigma}_0^{-1} (f(x) - \hat{\mu}_0)$.

This formulation is mathematically equivalent to a log-likelihood ratio test between a class-conditional model and a background model.

$$RMD(x) \approx -\log P(x|\text{Class}_k) + \log P(x|\text{Background})$$

By subtracting the background term, RMD cancels out the common factors (e.g., "it is a leaf"). A high RMD score implies that the sample is close to the class centroid *relative* to how close it is to the generic background. If an image is just a generic leaf (Near-OOD), it will be close to μ_0 and moderately close to μ_c, resulting in a small difference. If it is a specific disease instance (ID), it will be very close to μ_c but generic relative to μ_0, maximizing the score.

2.2.3 Empirical Validation in Fine-Grained Tasks

The efficacy of RMD is most pronounced in fine-grained tasks. On the Genomics OOD benchmark—which involves distinguishing between DNA sequences from different bacterial classes (highly similar)—RMD improved the Area Under the Receiver Operating Characteristic (AUROC) by nearly **16%** compared to standard MD.[11] In the context of plant diseases, this suggests that RMD is essential for distinguishing between visually similar pathologies.

2.3 Mahalanobis++: Addressing Feature Norm Instability

While RMD addresses the semantic overlap, recent research has highlighted a structural instability in MD related to the statistical properties of ViT features. The study "Mahalanobis++" [13] identifies that the L2 norms of feature vectors in modern pre-trained models are often heavy-tailed and not normally distributed.

2.3.1 The Norm-Variance Correlation

In many models, there is a strong correlation between the norm of the feature vector $\|z\|_2$ and the prediction confidence. OOD samples often result in feature vectors with significantly smaller (or larger) norms than ID samples. However, the standard MD calculation allows the covariance matrix to be dominated by the directions of high variance, which are often just the directions of magnitude change. This creates "blind spots" where OOD samples with typical norms but atypical angles are not detected.[15]

2.3.2 The Normalization Solution

Mahalanobis++ proposes a preprocessing step: **L2-normalization** of the features before computing the Gaussian statistics.

$$z_{norm} = \frac{z}{\|z\|_2}$$

By projecting all features onto the unit hypersphere, the method eliminates the variance due

to magnitude. The MD then becomes purely a measure of angular distance (cosine similarity) scaled by the angular spread of the class cluster. Experimental results on 44 different models (including ViTs and ConvNeXts) showed that this normalization consistently improved OOD detection performance, reducing the False Positive Rate at 95% True Positive Rate (FPR95) by an average of 7% compared to the state-of-the-art ViM method.[14]

For Vision Transformers in agriculture, which often process images with variable background clutter (leading to variable feature energy), Mahalanobis++ provides a crucial stabilization mechanism, ensuring that the detection is based on the *pattern* of the disease, not the *contrast* or brightness of the image.

---

3. Extreme Value Theory and the Legacy of OpenMax

Before the dominance of geometric distance methods, the field of Open Set Recognition (OSR) was defined by **OpenMax**, an algorithm grounded in Extreme Value Theory (EVT). Understanding OpenMax is crucial for historical context and for recognizing the specific limitations that modern methods must overcome.

3.1 The Statistical Basis: Extreme Value Theory (EVT)

EVT is a branch of statistics used to model the risk of extreme deviations from the median of probability distributions (e.g., "100-year floods"). The central theorem (Fisher-Tippett-Gnedenko) states that the maximum of a sequence of independent and identically distributed random variables converges to one of three distributions: Gumbel, Fréchet, or **Weibull**.[16]

In the context of neural networks, the "scores" (activations) of correct classes are considered the "normal" distribution. The "scores" of outliers or unknown classes are expected to fall into the "tail" of this distribution. By modeling this tail, one can estimate the probability that a given input belongs to the set of "knowns" or if it is an "extreme" value belonging to the "unknown."

3.2 The OpenMax Algorithm Deep Dive

OpenMax was designed as a drop-in replacement for the Softmax layer. Its operation involves two phases: training (calibration) and inference.

3.2.1 Phase 1: Calibration (Weibull Fitting)

1. **Feature Extraction:** For each training sample x_i correctly classified as class c, the activation vector $v(x_i)$ from the penultimate layer is extracted.

2. **Centroid Calculation:** The Mean Activation Vector (MAV) μ_c is computed for each class.
3. **Distance Calculation:** The Euclidean distance between each sample and its class MAV is computed: $d_i = \|v(x_i) - \mu_c\|_2$.
4. **Tail Modeling:** For each class, the largest distances (e.g., the top 20 outliers within the class) are selected. A **Weibull distribution** is fitted to these distances. The Weibull PDF provides the probability of a distance being "too large" for that class.[16]

3.2.2 Phase 2: Inference

1. **Prediction:** Given a test sample x, the distances to the top-α predicted classes are computed.
2. **Rejection Probability:** The fitted Weibull models are used to calculate the probability ω_c that the sample is an outlier for class c.
3. **Score Revision:** The activation (logit) for class c is recalibrated: $\hat{v}_c = v_c(1 - \omega_c) + \dots$. Essentially, if the outlier probability is high, the activation is dampened.
4. **Unknown Class:** The "subtracted" probability mass is assigned to a new pseudo-class y_{K+1} (the "Unknown" class).
5. **Softmax:** A Softmax is applied over the $K + 1$ classes.

3.3 The Failure of OpenMax in Modern Architectures

While OpenMax represented a significant conceptual leap, empirical evidence suggests it struggles in the feature spaces of modern Deep Neural Networks (DNNs), particularly ViTs.

1. **The Curse of Dimensionality:** EVT relies on the concept of distance. In high-dimensional spaces (ViT-Base has 768 dimensions), the distribution of distances becomes concentrated, and the distinction between the "bulk" and the "tail" blurs. The assumptions required for the Fisher-Tippett theorem are often violated by the highly correlated, manifold-bound features of DNNs.[19]
2. **Hyperparameter Sensitivity:** The performance of OpenMax is notoriously sensitive to the "tail size" parameter (how many samples are used to fit the Weibull). In fine-grained tasks, the "outliers" of a class (e.g., a heavily infected leaf) might overlap with the "inliers" of another class (e.g., a different disease), making the Weibull fit unstable.[20]
3. **Logit vs. Latent:** OpenMax modifies logits based on latent distances. However, Mahalanobis Distance methods have shown that operating *purely* in the latent space (pre-logit) is more effective. The logits are already compressed representations that discard vital geometric information needed for outlier detection.[7]

Synthesis: For agricultural ViTs, OpenMax is largely considered a legacy method. While its probabilistic formalism is attractive, its empirical performance on Near-OOD tasks is consistently inferior to RMD and Mahalanobis++, which handle the geometric subtleties of fine-grained classes more robustly.[21]

---

4. The Foundation Model Era: DINOv2 and the OOD Landscape

The paradigm of training models from scratch (supervised learning) is being superseded by the use of **Foundation Models**—large-scale networks pre-trained on massive datasets using Self-Supervised Learning (SSL). **DINOv2** (Discriminative Self-supervised Learning) represents the state-of-the-art in this domain, offering feature spaces that are remarkably robust for OOD detection without any task-specific fine-tuning.

4.1 DINOv2: Architecture and Pre-training Signals

DINOv2 employs a student-teacher architecture trained with a combination of **DINO** (self-distillation) and **iBOT** (Masked Image Modeling) losses.[22]

- **DINO Loss:** Encourages the student network to match the teacher's output on different augmented views of the same image (global consistency).
- **iBOT Loss:** Forces the student to reconstruct masked patches of the image based on the visible context (local consistency).

This dual objective is critical for fine-grained plant disease tasks. The iBOT loss compels the model to understand local textures (e.g., "this green patch implies the neighboring patch should be green," or "this yellow halo implies a fungal center"). Unlike CLIP, which aligns images to text captions (often losing fine-grained visual details), DINOv2 is purely visual. It learns a feature space where "texture" and "shape" are equally represented, making it superior for distinguishing between visually similar diseases.[24]

4.2 The Linear Probe vs. Nearest Neighbor (k-NN) Debate

When adapting a foundation model like DINOv2 for OOD detection, a critical implementation choice arises: Should we train a Linear Probe (classifier head) or use Non-Parametric Nearest Neighbor (k-NN) retrieval?

4.2.1 Linear Probing: The Bottleneck Effect

Linear probing involves freezing the DINOv2 backbone and training a linear layer W on the labeled ID dataset.

- **Mechanism:** The linear layer attempts to find hyperplanes that separate the ID classes.

- **OOD Weakness:** The "TEMI" study [26] and extensive benchmarks [27] reveal that linear probing often degrades OOD detection performance. The linear projection simplifies the complex, high-dimensional manifold of the DINOv2 features into a lower-dimensional decision space. OOD samples that lie far from the ID manifold in the *feature space* might project onto the "high confidence" side of the linear boundary in the *decision space*. This is the "feature collapse" phenomenon.

4.2.2 Nearest Neighbor (k-NN): Manifold Preservation

The k-NN approach stores the feature embeddings of the entire training set (the "Memory Bank"). For a test sample x, the OOD score is the distance to the k-th nearest neighbor in the bank.

- **Mechanism:** Non-parametric density estimation.
- **OOD Strength:** k-NN operates directly on the native DINOv2 manifold. Because DINOv2 is trained to cluster visually similar images, ID samples naturally cluster tightly. OOD samples, lacking the specific visual correlations learned during pre-training, fall into sparse regions of the space. The distance to the nearest neighbor is a high-fidelity proxy for $P(x)$.
- **Empirical Evidence:** In few-shot and zero-shot scenarios, k-NN detectors on DINOv2 features consistently outperform linear probes. For instance, on ImageNet benchmarks, k-NN maintained high AUROC even when the linear probe's accuracy dropped due to limited data.[27]

4.2.3 Implications for Agriculture

In agriculture, obtaining labeled data for every possible disease variant is impossible (the "Long Tail" problem). DINOv2 + k-NN allows for a flexible, evolving system. When a new disease appears, one can simply add the embeddings of a few verified samples to the memory bank without re-training a classifier. This "retrieval-based" classification is inherently more robust to the open-world nature of the field.[24]

---

5. Parameter-Efficient Fine-Tuning (LoRA) as an OOD Sensor

While DINOv2 provides excellent general features, detecting specific crop diseases often requires fine-tuning. However, full fine-tuning is computationally expensive and risks "catastrophic forgetting" of the general knowledge that aids OOD detection. **Low-Rank Adaptation (LoRA)** offers a solution that not only enables efficient adaptation but also introduces a novel mechanism for OOD detection.

5.1 LoRA Mathematics

LoRA hypothesizes that the change in weights during fine-tuning has a low intrinsic rank. Instead of updating the full weight matrix $W \in \mathbb{R}^{d \times k}$, LoRA injects two trainable low-rank matrices $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{k \times r}$, where $r \ll d$ (e.g., $r = 16$).

The forward pass for a layer becomes:

$$h = W_0 x + \Delta W x = W_0 x + \alpha B A x$$

where α is a scaling factor. Typically, for deployment, ΔW is merged into W_0 ($W' = W_0 + \alpha B A$) to eliminate the inference overhead of the separate branch.[29]

5.2 The "Beyond Fine-Tuning" Insight: Unmerged Embeddings

A breakthrough study titled *"Beyond Fine-Tuning: LoRA Modules Boost Near-OOD Detection"* [30] challenges the standard merging practice. The authors demonstrate that the **activations of the LoRA branch (BAx)** contain highly specific information about the ID task that is lost when merged.

5.2.1 The Adaptation Signal Hypothesis

The LoRA parameters A and B are trained *solely* on the ID dataset. Therefore, the term BAx represents the "adaptation signal"—the specific features required to process the ID data that were missing from the pre-trained backbone.

- **ID Sample:** The input x contains features relevant to the fine-tuned task (e.g., disease lesions). The adapter BAx activates strongly and coherently to process these features.
- **OOD Sample:** The input x lacks the specific features the adapter was trained to recognize. The adapter activation BAx is likely to be weak, random, or structurally distinct from ID activations.

5.3 LoRA-Based Mahalanobis Distance (LoRA-MD)

The proposed method extracts a specialized embedding vector $E_{LoRA}(x)$ by concatenating the adapter outputs across all layers:

$$E_{LoRA}(x) = \text{Concat}_{l=1}^{L}$$

The Mahalanobis Distance is then computed on these E_{LoRA} vectors rather than the backbone features.

5.3.1 Performance and Sensitivity

This method acts as a "Task-Specific Filter." By ignoring the backbone features (which might activate for any leaf) and focusing on the adapter features (which only activate for *specific* diseases), LoRA-MD drastically improves sensitivity to Near-OOD samples.

- **Empirical Results:** In fine-grained medical benchmarks (MedMCQA), LoRA-MD achieved an AUROC of **0.890**, compared to **0.428** for MD on the last layer of the base model.[30]
- **Model Versioning:** The sensitivity is so acute that LoRA-MD can distinguish between different versions of a model fine-tuned on slightly different datasets, providing a security mechanism for model lineage verification.[30]

Recommendation for Plant Disease: Fine-tune a ViT using LoRA (Rank 16-32) on the crop disease dataset. During inference, keep the adapters unmerged. Use the concatenated adapter activations to compute the Mahalanobis Distance. This effectively creates a detector that measures "How much does the model recognize this as a *disease*?" rather than "How much does it recognize this as a *leaf*?"

---

6. Fine-Grained Attention Mechanisms

The ability of a ViT to detect OOD samples is intrinsically linked to *what* it attends to. In FGVC, attention mechanisms must be specialized to capture multi-scale features.

6.1 Multi-Scale Attention Architectures

Standard ViT attention is global and single-scale. However, plant diseases manifest at disparate scales: tiny early-stage spots vs. large late-stage blights. **BiFormer** and **Cascaded Multi-Scale Attention (CMSA)** [31] address this by integrating attention across scales.

- **CMSA:** Fuses tokens from shallow layers (high spatial resolution) with deep layers (high semantic resolution). This ensures that the OOD detector has access to both the "texture of the spot" (shallow) and the "shape of the lesion" (deep).
- **Impact on OOD:** If an OOD sample (e.g., a synthetic fake leaf) has the correct global shape but the wrong local texture, a single-scale ViT might be fooled. A multi-scale attention network will detect the discrepancy between the shallow and deep features, leading to a higher OOD score (e.g., via Mahalanobis distance computed on concatenated multi-scale features).[33]

6.2 Attention Entropy as an Uncertainty Metric

The entropy of the attention map A provides a lightweight, training-free OOD signal.

$$H(A) = -\sum_{j} A_{ij} \log A_{ij}$$

- **ID Behavior:** For a known disease, the attention heads should sharply focus on the lesion. The attention distribution is "peaked" (Low Entropy).
- **OOD Behavior:** For an unknown object or disease, the attention mechanism often fails to find a salient anchor. The attention becomes diffuse or uniformly distributed across the background. (High Entropy).[34]

VIPAMIN: Recent work on prompt tuning [35] shows that optimizing prompts to minimize attention entropy on ID data widens the gap between ID and OOD entropy, making this heuristic even more effective.

---

7. Interpretability and Validation: Class Activation Mapping (CAM)

Trust is the currency of agricultural AI. A farmer will not act on a "Spray Fungicide" recommendation if the model cannot justify it. **Class Activation Mapping (CAM)** provides the visual verification layer for OOD detection.

7.1 Adapting Grad-CAM for ViTs

Traditional Grad-CAM relies on convolutional feature maps. For ViTs, we use **Attention Rollout** or **Grad-CAM applied to reshaped tokens**. The gradient of the target class score y^c with respect to the attention maps $A^{(k)}$ of the last layer is computed. The importance weights α_k^c are derived, and the weighted sum of attention maps produces the CAM.[4]

$$L_{CAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^{(k)}\right)$$

7.2 Diagnosing "Clever Hans" OOD Failures

OOD detectors are not perfect. They can produce False Negatives (classifying an OOD sample as ID). CAM serves as a final "sanity check."

- **Scenario:** A model classifies an image of a generic green surface (OOD) as "Healthy

Soybean" with high confidence (and low Mahalanobis distance due to background dominance).
- **CAM Diagnosis:** The CAM visualization shows the model attending to the corners of the image or random noise rather than a leaf structure.
- **Action:** The system flags this as "Low Reliability" despite the high OOD score, triggering human review. This visual validation loop is critical for preventing "Clever Hans" behaviors where models rely on spurious background correlations (e.g., detecting the soil type instead of the plant).[37]

---

8. Emerging Alternatives and Synthesis

Beyond the core methods discussed, **Energy-Based Models (EBM)** offer a powerful alternative. The Energy score $E(x) = -T \log \sum_c e^{f_c(x)/T}$ is theoretically aligned with the probability density $p(x)$.[39]

- **Comparison:** While simple to implement, EBMs often underperform Mahalanobis-based methods on *fine-grained* tasks because the logits (from which energy is derived) are often "confused" (high entropy) even for ID samples in FGVC, blurring the distinction with OOD. However, EBMs are excellent for detecting far-OOD samples (e.g., non-plant images).

8.1 Comparative Synthesis Table

The following table summarizes the suitability of each method for the specific constraints of Fine-Grained Plant Disease Detection.

Methodology	**Primary Mechanism**	**Near-OOD (Fine-Grained)**	**Far-OOD (Non-Plant)**	**Computational Cost**	**Implementation Complexity**	**Best For...**
Standard MD	Latent Gaussian Distance	Low (Background confounding)	High	Low	Low	General filtering
Relative MD (RMD)	Ratio (Class vs. Backgrou	**Very High**	High	Moderate (2x fit)	Moderate	**Distinguishing similar**

	nd)					**diseases**
Mahalanobis++	L2-Norm + MD	High (Stabilizes ViT)	High	Low	Low	**ViT Architectures**
OpenMax	EVT / Weibull Tail	Low (Curse of dim.)	Moderate	Moderate	High (Tuning)	Legacy comparisons
DINOv2 + k-NN	Non-parametric Density	**Very High**	High	High (Vector Search)	Moderate	**Few-Shot / Emerging Diseases**
LoRA-MD	Adapter "Surprise"	**Highest**	High	Moderate	High	**Fine-Tuned Specific Models**

---

9. Conclusion

The landscape of Out-of-Distribution detection for agricultural Vision Transformers is evolving rapidly. The limitations of classical EVT-based methods (OpenMax) in high-dimensional spaces have necessitated a shift toward geometric approaches.

For the specific challenge of **Fine-Grained Plant Disease Detection**, where the line between "Healthy" and "Diseased" is razor-thin, relying on global feature distances is insufficient. The most robust solution involves a composite framework:

1. **Architecture:** Utilize **DINOv2** backbones to leverage superior texture modeling.
2. **Adaptation:** Employ **LoRA** for task-specific fine-tuning without destroying the OOD-robust manifold.
3. **Detection:** Implement **Mahalanobis Distance on Unmerged LoRA Embeddings** (LoRA-MD) to capture the specific adaptation signal, augmented by **Relative Mahalanobis Distance (RMD)** to filter background noise.
4. **Verification:** Deploy **Grad-CAM** and **Attention Entropy** checks to ensure the decision is grounded in relevant pathological features.

This multi-layered approach addresses the theoretical vulnerabilities of Softmax confidence

while providing the practical robustness required for real-world agricultural deployment.

References (Integrated)

- [8] Lee, K., et al. (2018). *A Simple Unified Framework for Detecting Out-of-Distribution Samples...*. NeurIPS.
- [13] Mueller, M. & Hein, M. (2025). *Mahalanobis++: Improving OOD Detection via Feature Normalization*. ICML.
- [11] Ren, J., et al. (2021). *A Simple Fix to Mahalanobis Distance...*. arXiv.
- [16] Bendale, A. & Boult, T. (2016). *Towards Open Set Deep Networks* (OpenMax). CVPR.
- [24] Oquab, M., et al. (2023). *DINOv2: Learning Robust Visual Features without Supervision*.
- [30] Salimbeni, E., et al. (2024). *Beyond Fine-Tuning: LoRA Modules Boost Near-OOD Detection...*. IEEE DLSP.
- [1] Various authors on FGVC Plant Disease (2024-2025).
- [27] Comparison of Linear Probe vs k-NN for OOD.
- [34] Attention Entropy and Prompt Tuning for OOD.

Alıntılanan çalışmalar

1. IS-ViT: A Novel Model for Fine-Grained Visual Recognition Based on Image Segmentation - IEEE Xplore, erişim tarihi Şubat 2, 2026, https://ieeexplore.ieee.org/document/10695724/
2. Dual-Dependency Attention Transformer for Fine-Grained Visual Classification - PMC, erişim tarihi Şubat 2, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC11014298/
3. [2106.10587] Exploring Vision Transformers for Fine-grained Classification - arXiv, erişim tarihi Şubat 2, 2026, https://arxiv.org/abs/2106.10587
4. Enhancing multiclass plant disease classification using GAN-boosted vision transformer with XAI insights - NIH, erişim tarihi Şubat 2, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12827657/
5. Delving into Out-of-Distribution Detection with Vision-Language Representations - NeurIPS, erişim tarihi Şubat 2, 2026, https://papers.neurips.cc/paper_files/paper/2022/file/e43a33994a28f746dcfd53eb

51ed3c2d-Paper-Conference.pdf
6. Exploring the Limits of Out-of-Distribution Detection - NeurIPS, erişim tarihi Şubat 2, 2026, https://proceedings.neurips.cc/paper/2021/file/3941c4358616274ac2436eacf67fae05-Paper.pdf
7. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection - arXiv, erişim tarihi Şubat 2, 2026, https://arxiv.org/html/2306.09301v5
8. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks - NIPS, erişim tarihi Şubat 2, 2026, http://papers.neurips.cc/paper/7947-a-simple-unified-framework-for-detecting-out-of-distribution-samples-and-adversarial-attacks.pdf
9. A framework for studying the best practises for Mahalanobis distance for OOD detection - GitHub, erişim tarihi Şubat 2, 2026, https://github.com/HarryAnthony/Mahalanobis-OOD-detection
10. [2309.01488] On the use of Mahalanobis distance for out-of-distribution detection with neural networks for medical imaging - arXiv, erişim tarihi Şubat 2, 2026, https://arxiv.org/abs/2309.01488
11. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection - ResearchGate, erişim tarihi Şubat 2, 2026, https://www.researchgate.net/publication/352506355_A_Simple_Fix_to_Mahalanobis_Distance_for_Improving_Near-OOD_Detection
12. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection, erişim tarihi Şubat 2, 2026, https://www.gatsby.ucl.ac.uk/~balaji/udl2021/accepted-papers/UDL2021-paper-007.pdf
13. Mahalanobis++: Improving OOD Detection via Feature Normalization - arXiv, erişim tarihi Şubat 2, 2026, https://arxiv.org/html/2505.18032v1
14. ICML Poster Mahalanobis++: Improving OOD Detection via Feature Normalization, erişim tarihi Şubat 2, 2026, https://icml.cc/virtual/2025/poster/43649
15. Mahalanobis++: Improving OOD Detection via Feature Normalization - GitHub, erişim tarihi Şubat 2, 2026, https://raw.githubusercontent.com/mlresearch/v267/main/assets/muller25a/muller25a.pdf
16. A Dual-threshold Based Evidential Openmax Approach for Open Set Recognition - Onera, erişim tarihi Şubat 2, 2026, https://onera.fr/sites/default/files/297/Fusion%202024%20-%20Evidential%20openmax%20approach%20-%20postprint.pdf
17. Towards Open Set Deep Networks, erişim tarihi Şubat 2, 2026, https://openaccess.thecvf.com/content_cvpr_2016/papers/Bendale_Towards_Open_Set_CVPR_2016_paper.pdf
18. Generative OpenMax for Multi-Class Open Set Classification - BMVA Archive, erişim tarihi Şubat 2, 2026, https://bmva-archive.org.uk/bmvc/2017/papers/paper042/paper042.pdf
19. On Validity of Extreme Value Theory-Based Parametric Models for Out-of-Distribution Detection, erişim tarihi Şubat 2, 2026,

https://www.iccs-meeting.org/archive/iccs2021/papers/127440140.pdf
20. OpenMax with Clustering for Open-Set Classification - IFI UZH, erişim tarihi Şubat 2, 2026, https://www.ifi.uzh.ch/dam/jcr:85dfa060-839a-4d4c-83a0-58d334ee064c/huber2024bachelor.pdf
21. Combining pre-trained Vision Transformers and CIDER for Out Of Domain Detection - arXiv, erişim tarihi Şubat 2, 2026, https://arxiv.org/pdf/2309.03047
22. Paper Review: DINOv2: Learning Robust Visual Features without Supervision, erişim tarihi Şubat 2, 2026, https://andlukyane.com/blog/paper-review-dinov2
23. DINOv2: Learning Robust Visual Features without Supervision, erişim tarihi Şubat 2, 2026, https://arxiv.org/pdf/2304.07193
24. What you can use DINOv2 for (with practical Python examples) | by Analystmachinelearning, erişim tarihi Şubat 2, 2026, https://medium.com/@analystmachinelearning/what-you-can-use-dinov2-for-with-practical-python-examples-5d7979e3fb93
25. Rethinking Out-of-Distribution Detection in Vision Foundation Models - OpenReview, erişim tarihi Şubat 2, 2026, https://openreview.net/forum?id=awReGYZaGl
26. Unsupervised Image Classification with Adaptive Nearest Neighbor Selection and Cluster Ensembles - arXiv, erişim tarihi Şubat 2, 2026, https://arxiv.org/html/2511.16213v1
27. Enhancing the Power of OOD Detection via Sample-Aware Model Selection - CVF Open Access, erişim tarihi Şubat 2, 2026, https://openaccess.thecvf.com/content/CVPR2024/papers/Xue_Enhancing_the_Power_of_OOD_Detection_via_Sample-Aware_Model_Selection_CVPR_2024_paper.pdf
28. Towards Few-shot Out-of-Distribution Detection - arXiv, erişim tarihi Şubat 2, 2026, https://arxiv.org/html/2311.12076v3
29. LoRA - Hugging Face, erişim tarihi Şubat 2, 2026, https://huggingface.co/docs/peft/en/package_reference/lora
30. Beyond fine-tuning: LoRA modules boost near-OOD ... - DLSP 2024, erişim tarihi Şubat 2, 2026, https://dlsp2024.ieee-security.org/papers/dls2024-final19.pdf
31. Multi-Scale Attention Networks with Feature Refinement for Medical Item Classification in Intelligent Healthcare Systems - NIH, erişim tarihi Şubat 2, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12431172/
32. Cascaded Multi-Scale Attention for Enhanced Multi-Scale Feature Extraction and Interaction with Low-Resolution Images - arXiv, erişim tarihi Şubat 2, 2026, https://arxiv.org/html/2412.02197v1
33. Multi-Scale Attention-Driven Hierarchical Learning for Fine-Grained Visual Categorization, erişim tarihi Şubat 2, 2026, https://www.mdpi.com/2079-9292/14/14/2869
34. VIPAMIN: Visual Prompt Initialization via Embedding Selection and Subspace Expansion - OpenReview, erişim tarihi Şubat 2, 2026, https://openreview.net/pdf?id=o0AASFieVc
35. VIPAMIN: Visual Prompt Initialization via Embedding Selection and Subspace

Expansion, erişim tarihi Şubat 2, 2026, https://arxiv.org/html/2510.16446v1
36. [Project] Recent Class Activation Map Methods for CNNs and Vision Transformers - Reddit, erişim tarihi Şubat 2, 2026, https://www.reddit.com/r/MachineLearning/comments/myenmh/project_recent_class_activation_map_methods_for/
37. Enhanced plant disease classification with attention-based convolutional neural network using squeeze and excitation mechanism - PubMed Central, erişim tarihi Şubat 2, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12378314/
38. Implementation of Explainable AI in Deep Learning Methods for Multiclass Classification of Plant Diseases in Mango Leaves - ddd-UAB, erişim tarihi Şubat 2, 2026, https://ddd.uab.cat/pub/elcvia/elcvia_a2025v24n1/elcvia_a2025v24n1p104.pdf
39. The energy-based OOD detection framework. The energy function maps the... - ResearchGate, erişim tarihi Şubat 2, 2026, https://www.researchgate.net/figure/The-energy-based-OOD-detection-framework-The-energy-function-maps-the-logit-outputs-to-a_fig11_355698553
40. Energy-based Out-of-distribution Detection - NIPS - NeurIPS, erişim tarihi Şubat 2, 2026, https://proceedings.neurips.cc/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf