Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Resource-Efficient Few-Shot Plant Disease Classification via Quantized Low-Rank Adapters in Vision Transformers

Seda BAYAT TOKSOZ

bayatseda@gmail.com

Gültekin ışık

Research Article

Keywords: Few-Shot Learning, Plant Disease Classification, Vision Transformers, Quantized Low-Rank Adaptation (QLoRA), Digital Agriculture

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9014055/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Resource-Efficient Few-Shot Plant Disease Classification via Quantized Low-Rank Adapters in Vision Transformers

Gültekin Işık[1*], Seda Bayat Toksöz[1*], Fesih Keskin[1*]

1: Igdir University, Computer Engineering Department, Igdir, Türkiye

*: corresponding author, {gultekin.isik, seda.bayat, fesih.keskin}@igdir.edu.tr

(Orcid ID: 0000-0003-3037-5586, 0000-0002-8427-9971, 0000-0002-3798-2912)

Abstract

Rapid and accurate detection of plant diseases from limited labeled data remains a critical challenge in digital agriculture. In this paper, we present a few-shot learning framework for plant disease classification that integrates a self-supervised Vision Transformer backbone (DINOv2-S) with a Prototypical Network classifier, adapted via Quantized Low-Rank Adaptation (QLoRA). While QLoRA has demonstrated strong efficiency gains in natural language processing since its introduction, its application to vision-domain few-shot learning tasks has not been systematically explored. Our framework fine-tunes only approximately 1% of the model parameters (~221K) using low-rank adapters and 4-bit quantization, achieving an approximate trainable-parameter storage of 2.53 MB and an inference speed of 0.20 ms/image on NVIDIA A100 hardware. Comprehensive experiments on the PlantDoc and PlantVillage datasets show competitive or improved performance relative to recent approaches. In the 5-way 5-shot setting at 224×224 resolution, the proposed framework achieves mean accuracies of 86.85 ± 0.42% and 94.51 ± 0.35%, respectively. Furthermore, ablation studies across varying shot numbers, image resolutions, backbone architectures, LoRA rank configurations, and fine-tuning strategies confirm the robustness and efficiency of the proposed approach. These results suggest that parameter-efficient adaptation of vision transformers offers a practical pathway for deploying disease diagnosis systems in resource-constrained agricultural settings.

Keywords: Few-Shot Learning, Plant Disease Classification, Vision Transformers, Quantized Low-Rank Adaptation (QLoRA), Digital Agriculture.

1. Introduction

Plant diseases pose a significant threat to global agriculture, causing an estimated 20–40% annual yield losses, economic damage, and food insecurity, particularly in developing regions (Richard et al., 2022). Traditional detection methods, including visual inspection and laboratory analyses, are labor-intensive, slow, and poorly equipped to address emerging pathogens driven by climate change and global trade (Rezaei et al., 2024). Rapid, scalable, and adaptable detection systems are therefore essential for early intervention and sustainable farming practices.

Deep learning has substantially advanced plant disease detection by automating the identification of diseases from images with high accuracy and speed (Ferentinos, 2018). Convolutional neural networks (CNNs) have been widely adopted for their ability to extract hierarchical features that enable differentiation of subtle visual cues across disease categories (Gündüz & Işık, 2023). However, the success of deep learning models depends heavily on the availability of large, annotated datasets, which are often difficult to assemble in agricultural contexts due to cost and logistical constraints (Fan et al., 2022; Uzhinskiy, 2025). This challenge is particularly acute for rare or newly emerging diseases, where only a handful of labeled examples may exist. Data scarcity limits model performance and hinders generalization across diverse plant species (Işık, 2023), environmental conditions, and disease manifestations (Günde et al., 2026; Işık & Paçal, 2024).

Few-shot learning (FSL) has emerged as a promising approach to address data scarcity in plant disease recognition (Sun et al., 2024). By leveraging meta-learning techniques, FSL enables models to rapidly adapt to new disease classes from a small number of examples (Yang et al., 2024). Metric-based meta-learning approaches, such as Prototypical Networks (Snell et al., 2017), have proven particularly effective in this context. These methods learn an embedding space where each class is represented by a prototype, and classification is performed by measuring the distance

between query images and these prototypes. Rezaei et al. (2024) demonstrated that a Vision Transformer-based FSL system with feature-level attention could achieve over 94% accuracy on PlantVillage using five training images per class.

The integration of Vision Transformers (ViTs) has further advanced few-shot plant disease classification by addressing limitations inherent in CNNs. Unlike CNNs, which rely on local receptive fields and are constrained in capturing long-range dependencies, ViTs employ self-attention mechanisms to model global relationships within images (Dosovitskiy et al., 2020). This capability is beneficial for plant disease images, which often feature complex backgrounds, occlusions, or spatially distributed symptoms. Pre-trained foundation models such as DINOv2 (Oquab et al., 2023), trained via self-supervised learning on large-scale diverse datasets, provide generalizable feature representations that can be adapted to downstream tasks with minimal fine-tuning. However, adapting these large models to small datasets poses challenges, including the risk of overfitting and the computational costs associated with full model fine-tuning.

Parameter-efficient fine-tuning (PEFT) techniques have been developed to address these challenges, enabling adaptation of large models by updating only a small fraction of parameters. Low-Rank Adaptation (LoRA; Hu et al., 2022) decomposes weight updates into low-rank matrices, substantially reducing memory requirements. Its quantized variant, QLoRA (Dettmers et al., 2023), further combines low-rank updates with 4-bit quantization, enabling fine-tuning of large models on modest hardware. Since its introduction, QLoRA has been primarily applied and validated in natural language processing tasks. Its systematic application to computer vision, and specifically to few-shot learning with Vision Transformers, remains largely unexplored.

Several alternative PEFT methods exist for adapting vision models, including visual prompt tuning (VPT; Jia et al., 2022), which prepends learnable tokens to the input sequence, and adapter modules (Houlsby et al., 2019), which insert small bottleneck layers within transformer blocks. BitFit (Zaken et al., 2022) offers an even more constrained approach by updating only bias terms. Each method represents a different trade-off between adaptation capacity, memory efficiency, and computational cost. A key question is whether the combination of low-rank decomposition with aggressive quantization, as implemented in QLoRA, provides advantages over these alternatives in the few-shot vision setting.

In this work, we propose a few-shot learning framework for plant disease classification that integrates DINOv2-S as a visual backbone, QLoRA for parameter-efficient adaptation, and Prototypical Networks for metric-based classification. The specific contributions of this study are as follows:

(1) We present a systematic empirical evaluation of QLoRA-based adaptation for Vision Transformers in few-shot plant disease classification, demonstrating its viability beyond the natural language processing settings in which QLoRA was originally developed and validated.

(2) We integrate DINOv2-S with QLoRA adapters and a Prototypical Network classifier in a unified episodic meta-learning framework, achieving competitive accuracy on the PlantVillage and PlantDoc benchmarks while fine-tuning only ~1% of model parameters with an approximate trainable-parameter storage of 2.53 MB.

(3) We provide comprehensive ablation studies examining the effects of shot number, image resolution, backbone architecture, LoRA rank, and fine-tuning strategy, and compare QLoRA against standard LoRA, Full Fine-Tuning, and Partial Fine-Tuning to isolate the contributions of quantization and low-rank adaptation.

(4) We evaluate the practical efficiency of the proposed framework, reporting trainable parameter counts, approximate trainable-parameter storage, peak GPU memory during training, and inference latency, providing a concrete basis for assessing deployment feasibility in resource-constrained settings.

2. Related Works

2.1 Few-Shot Learning for Plant Disease Recognition

Few-shot learning has gained significant attention as a strategy for addressing classification tasks with minimal labeled data, a challenge particularly relevant in agricultural applications (Chen et al., 2021; Fan et al., 2022). Early FSL methods, including Matching Networks (Vinyals et al., 2016), Siamese Networks (Koch et al., 2015), and Prototypical Networks (Snell et al., 2017), introduced meta-learning paradigms that enable rapid adaptation to new tasks from limited examples. These foundational approaches have been progressively refined to improve generalization and robustness in agricultural contexts (de Andrade Porto et al., 2023; Wu et al., 2023).

Subsequent work has expanded the applicability of FSL to plant disease detection through integration of sophisticated meta-learning strategies. Chen et al. (2021) proposed LFM-CNAPS, a framework employing local feature matching to identify unseen disease classes with limited training data. Mu et al. (2024) combined supervised contrastive pre-training with prototypical meta-learning, achieving 79.5% accuracy on real-world plant disease datasets using only 30 labeled images. Wu et al. (2023) demonstrated through systematic evaluation that meta-learning approaches consistently outperform standard transfer learning when fewer than 20 labeled samples are available per class. These developments highlight the potential of FSL to address data scarcity, though many earlier studies relied on CNN backbones that may struggle to capture complex contextual relationships in plant disease images, particularly those with diverse or cluttered backgrounds (Ullah et al., 2022).

2.2 Vision Transformers in Few-Shot Classification

The advent of Vision Transformers has marked a significant shift in image recognition, including plant disease classification (Dosovitskiy et al., 2020). Pre-trained ViT models such as DINOv2 (Oquab et al., 2023) provide robust feature representations derived from extensive self-supervised training, enabling effective transfer learning with minimal fine-tuning. Rezaei et al. (2024) demonstrated the effectiveness of ViT-based FSL systems incorporating feature attention modules, outperforming CNN-based methods on challenging datasets such as PlantDoc. Metric-based meta-learning approaches, particularly Prototypical Networks (Snell et al., 2017), have remained a cornerstone of FSL due to their simplicity and effectiveness. Recent innovations such as PrototypeFormer (Su et al., 2025) and cross-attention techniques (Hiller et al., 2022) have further refined prototype-based classification.

2.3 Parameter-Efficient Fine-Tuning

Efficient adaptation of large pre-trained models to specific tasks with limited data is a critical challenge for deploying FSL systems. Full fine-tuning of all model parameters often leads to overfitting and demands substantial computational resources (Fan et al., 2022). Parameter-efficient fine-tuning methods have emerged as practical alternatives. LoRA (Hu et al., 2022) introduces low-rank decomposition of weight updates, enabling adaptation with a small subset of parameters. QLoRA (Dettmers et al., 2023) extends this approach with 4-bit quantization, further reducing storage requirements.

Several other PEFT methods have been proposed for vision transformers. Visual Prompt Tuning (VPT; Jia et al., 2022) prepends learnable prompt tokens to the input sequence, adapting the model behavior without modifying its weights. Adapter modules (Houlsby et al., 2019) insert lightweight bottleneck layers within transformer blocks, providing task-specific capacity with minimal parameter overhead. BitFit (Zaken et al., 2022) takes a minimalist approach by fine-tuning only bias terms. These methods have shown varying degrees of effectiveness depending on the task, dataset scale, and backbone architecture.

While QLoRA has seen widespread adoption in natural language processing, its application to vision tasks, particularly in the few-shot learning context, remains limited. Our work addresses this gap by systematically evaluating QLoRA-based adaptation of a DINOv2 Vision Transformer for few-shot plant disease recognition, providing comparisons against both standard LoRA (without quantization) and alternative fine-tuning strategies to isolate the contribution of each design choice.

3. Methodology

Our methodology is designed to address three primary challenges in few-shot plant disease classification: (i) acquiring robust and transferable visual representations, (ii) efficiently adapting pre-trained models without overfitting, and (iii) generalizing to novel classes with minimal data. We propose a unified framework that integrates a self-supervised Vision Transformer backbone (DINOv2-S), Quantized Low-Rank Adaptation for parameter-efficient fine-tuning, and a Prototypical Network classifier optimized via episodic meta-learning.

3.1 Overview of the Proposed Framework

The proposed framework operates in three interconnected stages. First, input images are processed by the pre-trained DINOv2-S Vision Transformer, which is kept largely frozen and quantized to 4-bit precision to minimize memory usage (Dettmers et al., 2023). Second, lightweight low-rank adapters, introduced via QLoRA, are injected into the attention layers of the transformer, enabling task-specific adaptation with approximately 1% of total parameters being trainable. Third, the resulting embeddings are fed into a Prototypical Network, which computes class prototypes from a support set and classifies query images based on their Euclidean distance to these prototypes (Snell et al., 2017). The entire pipeline is trained end-to-end using episodic meta-learning, and at inference, classification is performed directly via prototype matching without additional optimization. Figure 1 illustrates the overall architecture.

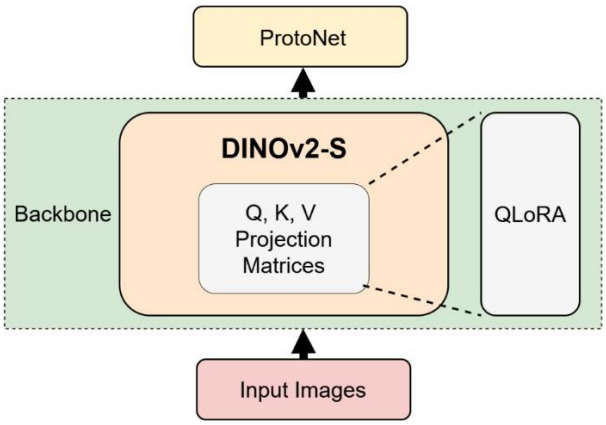


Figure 1. Overview of the proposed few-shot plant disease classification framework. Input images are first encoded by a 4-bit–quantized DINOv2-S backbone with low-rank QLoRA adapters injected into its attention layers; the resulting embeddings are then fed into a Prototypical Network to compute class prototypes and classify query samples in an N-way K-shot episodic setup, with direct prototype matching at inference.

3.2 Vision Transformer Backbone: DINOv2-S

We adopt DINOv2-S (Oquab et al., 2023), a compact variant of the ViT architecture, as the backbone for feature extraction. DINOv2-S comprises 12 transformer blocks, each with 384 hidden dimensions and 6 attention heads. The model processes images by dividing them into non-overlapping 14×14 pixel patches, which are linearly embedded into token sequences. These tokens are transformed through successive layers of multi-head self-attention and feed-forward networks, yielding contextualized embeddings that capture both local and global image features.

The strength of DINOv2-S lies in its self-supervised pre-training via self-distillation on the LVD-142M dataset, a collection of 142 million diverse images (Oquab et al., 2023). This pre-training produces highly generalizable visual representations, reducing dependence on labeled data and enhancing suitability for few-shot tasks where annotated samples are scarce. The self-attention mechanism enables the model to capture long-range dependencies within images, which is relevant for detecting subtle, spatially distributed disease symptoms. By leveraging these pre-trained features, the framework aims to minimize overfitting and achieve robust performance across varied disease datasets.

3.3 QLoRA-Based Adaptation

To adapt the pre-trained DINOv2-S model to the plant disease classification task, we employ QLoRA, a technique that integrates LoRA with quantization for efficient fine-tuning (Dettmers et al., 2023). QLoRA introduces lightweight trainable adapters into the transformer while quantizing its pre-trained weights to 4-bit precision, reducing memory and computational demands. Low-rank adapter modules are injected into each attention layer's Q, K, V (Query, Key, Value) projections. Figure 2 illustrates the LoRA and QLoRA injection mechanism.

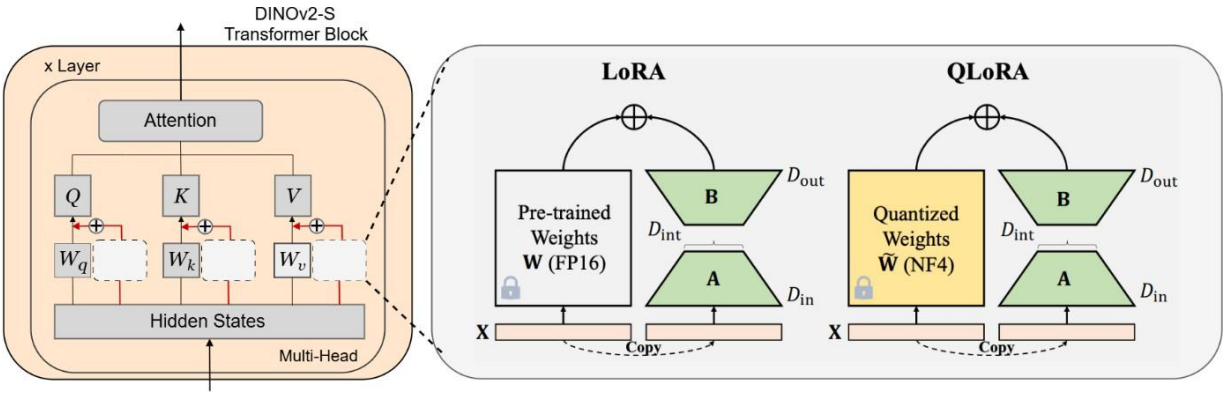


Figure 2. LoRA and QLoRA injection into DINOv2-S attention layers. In each transformer block, Q, K, V projection matrices W (left) are augmented by low-rank update factors A and B (LoRA), added to the original FP16 weights. Under QLoRA (right), the backbone weights are first quantized to NF4 (4-bit), then only the same low-rank adapters A, B remain in full precision, minimizing both trainable parameters and memory (Bayat Toksöz & Işık, 2025, 2026).

The LoRA component decomposes weight updates into two low-rank matrices, W_down and W_up, with the adapted weight matrix W' defined as:

$$W' = W + \left(\frac{\alpha}{r}\right) \cdot W_{up} \cdot W_{down} \quad (1)$$

where W represents the original frozen weights, r is the rank of the adaptation matrices, and α is a scaling factor. Based on preliminary experiments (see Section 5.5.4), we set r = 8 and α = 16 as the default configuration. A dropout rate of 0.05 is applied to the adapters for regularization. These adapters are inserted into the Q, K, V and output projection layers of each transformer block, targeting components most relevant for task-specific feature modulation.

This configuration reduces the trainable parameters to approximately 221K, less than 1% of DINOv2-S's total parameters (~21M), while the 4-bit quantization further reduces the storage requirements of the frozen backbone weights (from approximately 42 MB in FP16 to approximately 10.5 MB in NF4). QLoRA preserves the model's pre-trained knowledge and is expected to mitigate overfitting relative to approaches that update a larger fraction of parameters, making it a suitable adaptation method for the few-shot setting (Dettmers et al., 2023; Hu et al., 2022).

3.4 Prototypical Networks for Few-Shot Classification

For the classification stage, we utilize Prototypical Networks (ProtoNet), a metric-based meta-learning approach designed for few-shot learning (Snell et al., 2017). ProtoNet learns an embedding space where each class is represented by a prototype, computed as the mean of the embeddings of its support samples. Given an N-way K-shot support set $S = \{(x_i, y_i)\}$, the prototype c_n for class n is calculated as:

$$c_n = \left(\frac{1}{K}\right) \sum f_{\theta(x_i)}, for\ (x_i, y_i) \in S_n \quad (2)$$

where f_θ is the embedding function from the QLoRA-adapted DINOv2-S, and S_n is the subset of support samples for class n. For a query image x_q, classification is performed by assigning it to the nearest prototype based on squared Euclidean distance:

$$y_q * = \ argmin_n \left|\left| f_{\theta(x_q)} - \ c_n \right|\right|^2 \quad (3)$$

Training employs episodic meta-learning, with each episode comprising a support set and a query set. The model minimizes cross-entropy loss on the query set, optimizing the embedding space for compact, well-separated class clusters. At inference, new classes are classified directly via prototype matching without further optimization. Figure 3 illustrates the meta-learning scheme.

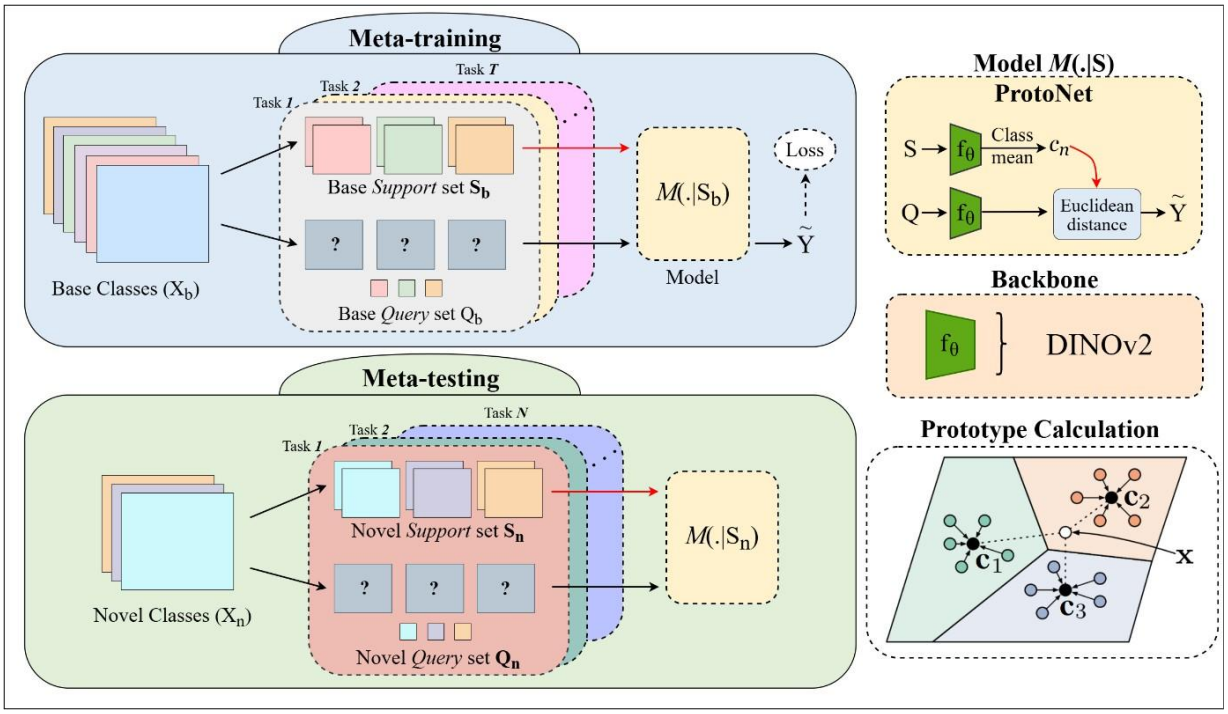


Figure 3. Meta-learning scheme for few-shot classification. During meta-training, tasks of N-way K-shot support sets S_b and query sets Q_b are sampled from base classes X_b and the model $M(\cdot)$ is optimized episodically. In meta-testing, novel classes X_n generate support S_n and query Q_n tasks to evaluate generalization. Right panel shows ProtoNet classification and query assignment by Euclidean distance. Support examples S and a query x are encoded by f_θ into embedding space; class prototypes c_n are formed as the mean of each class's support embeddings, and x is assigned to its nearest prototype by Euclidean. Figure adapted from (Günde & Işık, 2025; Işık & Paçal, 2024).

4. Experimental Setup

4.1 Datasets and Evaluation Protocol

4.1.1 Datasets

PlantDoc: This dataset includes 2,598 images distributed across 27 classes encompassing 13 plant species. Captured in real-world field conditions, the images exhibit varied backgrounds, inconsistent lighting, occlusions, and multi-scale resolutions. These attributes make PlantDoc a demanding testbed that reflects practical agricultural environments (Singh et al., 2020).

PlantVillage: Comprising 54,306 high-resolution images across 38 classes, PlantVillage offers a controlled laboratory setting with uniform backgrounds and distinct symptom visibility (Hughes et al., 2015; Mohanty et al., 2016). While less complex than PlantDoc, it serves as a widely adopted baseline for plant pathology research, facilitating comparisons with existing literature. Figure 4 shows representative images from both datasets.

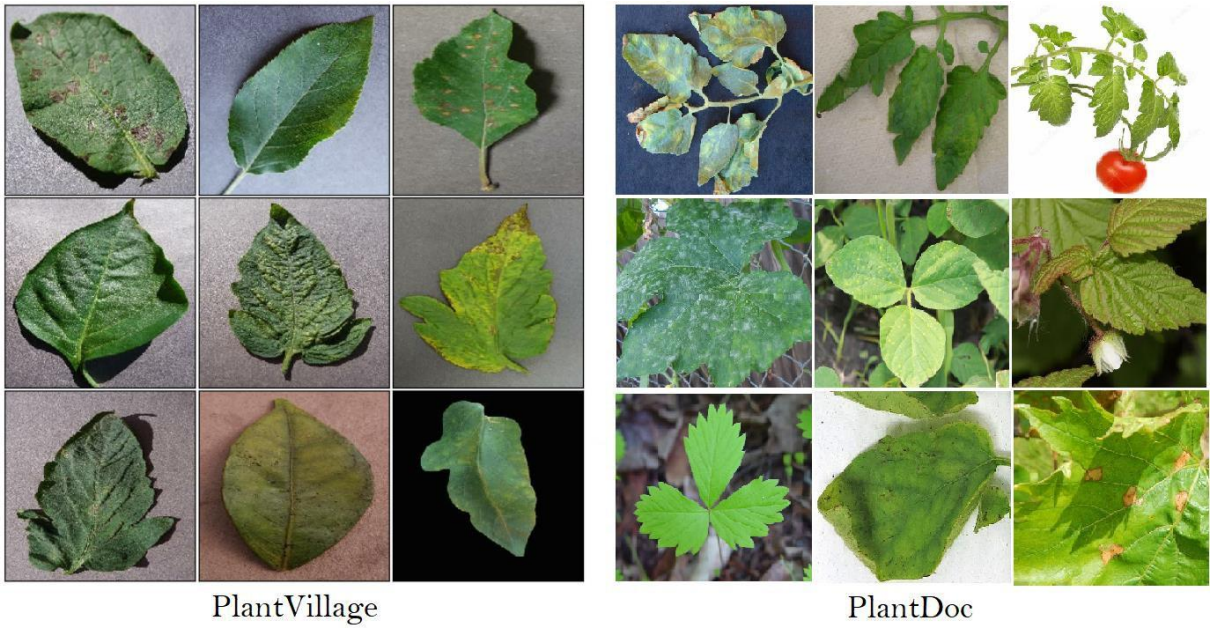


Figure 4. Representative images from PlantVillage (left) and PlantDoc (right), highlighting differences in complexity and image quality.

4.1.2 Data Split and Few-Shot Tasks

Following established few-shot learning conventions, the datasets are partitioned into training, validation, and test subsets with non-overlapping classes, maintaining an approximate split of 60% training, 20% validation, and 20% testing. Evaluation is structured as 5-way K-shot tasks, where K varies as 1, 5, 10, and 20, with 15 query samples per class.

For training, 10,000 episodes were sampled from the training classes. During validation, 1,000 episodes were drawn from the validation classes. For testing, three independent runs were conducted, each comprising 600 episodes from the test classes, totaling 1,800 episodes, to provide reliable estimates with 95% confidence intervals.

4.2 Implementation and Training Details

4.2.1 Data Preprocessing and Augmentation

To enhance model robustness, augmentation techniques applied during training include Random Resized Crop, Random Horizontal Flip with 50% probability, and Color Jitter to simulate environmental variations. By default, images are resized to 128×128 pixels. For ablation studies and comparisons with prior work, resolutions of 224×224 and 518×518 are also investigated.

4.2.2 Optimization Details

Training is optimized using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of $1e-4$ and gradient clipping with a maximum norm of 1.0. A weight decay of $1e-4$ is applied in the Partial Fine-Tuning configuration

to mitigate overfitting; for QLoRA and Full Fine-Tuning, no weight decay is used. This design choice was made because QLoRA's inherent parameter constraint provides implicit regularization, while Full Fine-Tuning was evaluated without weight decay to observe its unregularized behavior for comparison purposes. A Cosine Annealing learning rate scheduler is employed with $T_{max} = 10{,}000$ episodes.

4.2.3 Hardware and Computational Resources

All experiments are executed on NVIDIA A100 GPUs with 40 GB VRAM. Inference latency was measured using a batch size of 1, after 50 warm-up iterations, averaged over 500 timed forward passes with explicit CUDA synchronization. Reported values correspond to forward-pass inference excluding data loading and preprocessing. Performance on resource-constrained edge devices may differ substantially and is discussed as a direction for future work in Section 6.

4.3 Evaluation Metrics

Classification performance is evaluated through accuracy, precision, recall, and F1-score, measured on query images in the few-shot tasks. We report mean values and 95% confidence intervals across 1,800 test episodes (3 runs of 600 episodes each). Additionally, we report trainable parameter count, approximate trainable-parameter storage, and inference time per image to provide a comprehensive evaluation of both efficacy and efficiency.

4.4 Comparison Methods

We compare the following adaptation strategies to contextualize the contributions of QLoRA:

Full Fine-Tuning: Updates all 21.8M parameters of DINOv2-S, maximizing adaptability but increasing the risk of overfitting and resource demands.

Partial Fine-Tuning: Updates the final two transformer blocks, normalization layers, and classification head (3.6M parameters), providing a balance between adaptation and efficiency.

Standard LoRA (without quantization): Updates low-rank adapters with the same rank ($r = 8$) and scaling factor ($\alpha = 16$) as QLoRA, but with backbone weights maintained in FP16 precision. This comparison isolates the effect of 4-bit quantization.

QLoRA (proposed): Fine-tunes 221K parameters (~1%) using 4-bit quantized backbone weights and low-rank adapters in full precision.

Table 1 and Figure 5 summarize the computational characteristics of each strategy.

Table 1. Comparison of fine-tuning strategies for few-shot plant disease classification.

Method	Trainable Params	Trainable %	Model Precision	Trainable-Param. Storage	Inference Time
QLoRA	221k	≈ 1%	4-bit + fp16	≈ 2.5 MB	0.19-0.22 ms/img
Partial Fine-tune	3.6M	16.5%	fp32	≈ 42 MB	0.46 ms/img
Full Fine-tune	21.8M	100%	fp32	≈ 250 MB	1.12 ms/img

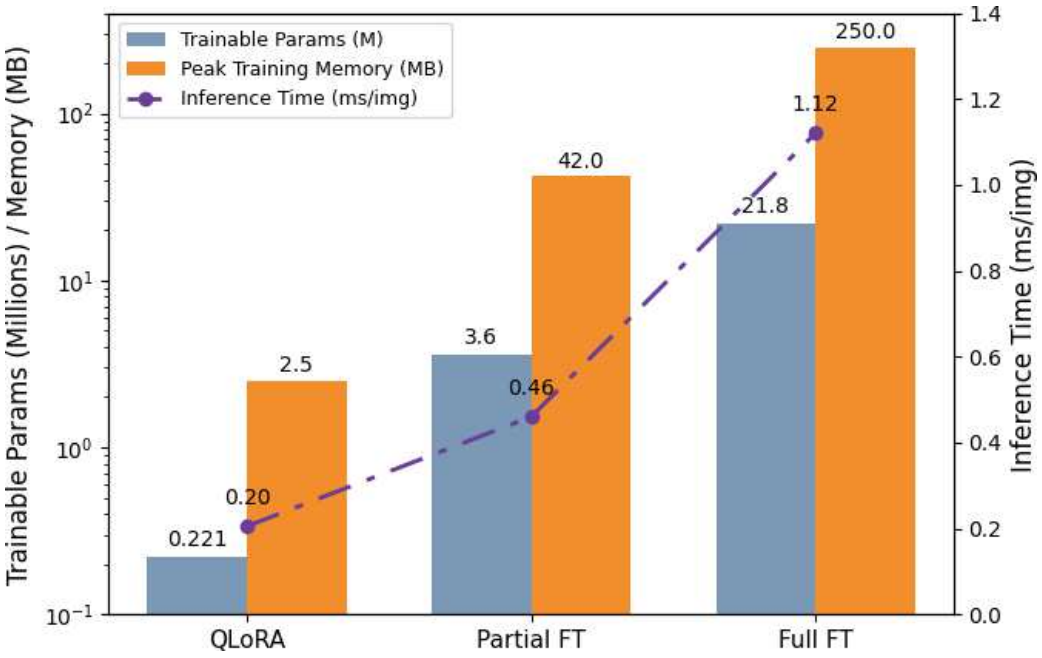


Figure 5. Comparison of fine-tuning strategies in terms of trainable parameter count (log-scale), approximate trainable-parameter storage (MB), and inference time (ms/image). QLoRA updates only ~1 % of the weights and requires ≈2.5 MB of trainable-parameter storage, whereas Partial and Full fine-tuning consume 42 MB and 250 MB respectively, with a corresponding rise in inference latency.

5. Results

This section presents classification results, ablation studies, comparisons with prior work, and feature space visualizations. All accuracy values are reported as mean ± 95% confidence interval across 1,800 test episodes unless otherwise noted.

5.1 Classification Performance

Table 2 presents results for both 3-way and 5-way classification tasks across varying shot settings.

Table 2. DINOv2-S + QLoRA + ProtoNet classification results (Image size = 128×128). Mean ± 95% CI over 1,800 test episodes.

Setting	PV Acc	PV Prec	PV Rec	PV F1	PD Acc	PD Prec	PD Rec	PD F1
3w 1s	87.15±0.48	88.49	87.15	86.72	80.12±0.55	81.05	80.12	79.76
3w 5s	94.30±0.32	95.25	94.35	93.86	87.25±0.46	88.21	87.21	86.99
3w 10s	95.76±0.28	96.20	95.82	95.49	91.86±0.38	92.45	91.84	91.55
3w 20s	98.55±0.15	98.78	98.61	98.05	92.36±0.35	92.89	92.45	92.24

5w 1s	83.27±0.52	84.15	83.31	82.77	75.91±0.61	76.22	76.02	75.56
5w 5s	91.45±0.38	91.85	91.45	91.12	83.69±0.49	83.78	83.86	83.55
5w 10s	93.96±0.31	94.19	94.01	93.16	87.15±0.44	87.62	87.28	86.82
5w 20s	96.19±0.24	96.22	96.28	95.98	88.36±0.41	88.89	88.35	88.29

On PlantVillage, accuracy improves consistently from 83.27 ± 0.52% (1-shot) to 96.19 ± 0.24% (20-shot) in the 5-way setting. The model achieves an F1-score of 91.12% in the 5-shot scenario, indicating balanced precision and recall. For the more challenging PlantDoc dataset, the framework achieves 83.69 ± 0.49% accuracy and 83.55% F1-score in the 5-way 5-shot task. The close alignment between Precision (83.78%) and Recall (83.86%) in this setting suggests that the model does not exhibit systematic bias toward particular classes.

5.2 Fine-Tuning Strategy Comparison

Table 3 compares QLoRA, standard LoRA (FP16 backbone), Partial Fine-Tuning, and Full Fine-Tuning across 1-shot, 5-shot, 10-shot, and 20-shot settings.

Table 3. Comparison of fine-tuning strategies across shot settings (5-way, 128×128). Mean accuracy (%) ± 95% confidence interval over 1,800 test episodes (3 runs, 600 tasks/run).

Setting	**QLoRA PV**	**LoRA PV**	**Part. PV**	**Full PV**	**QLoRA PD**	**LoRA PD**	**Part. PD**	**Full PD**
5w 1s	83.27±0.52	82.89±0.54	80.47±0.58	60.96±0.71	75.91±0.61	75.42±0.63	73.63±0.65	71.56±0.68
5w 5s	91.45±0.38	90.98±0.40	87.72±0.46	86.41±0.49	83.69±0.49	83.21±0.51	78.69±0.57	76.20±0.60
5w 10s	93.96±0.31	93.55±0.33	91.28±0.39	90.72±0.41	87.15±0.44	86.64±0.46	83.45±0.52	81.87±0.54
5w 20s	96.19±0.24	95.82±0.26	94.56±0.30	95.12±0.28	88.36±0.41	87.91±0.43	86.22±0.47	85.38±0.49

QLoRA consistently achieves the highest accuracy across all settings and both datasets. The advantage is most pronounced in low-shot scenarios: in the 5-way 1-shot setting, QLoRA outperforms Full Fine-Tuning by 22.3 percentage points on PlantVillage and 4.4 points on PlantDoc, reflecting the severity of overfitting when all parameters are updated with minimal data. Standard LoRA (without quantization) performs slightly below QLoRA, suggesting that 4-bit quantization does not degrade and may marginally improve performance by acting as an implicit regularizer on the frozen backbone. At higher shot counts (20-shot), the gap between methods narrows, as Full Fine-Tuning benefits from additional data; however, QLoRA maintains a consistent advantage while requiring approximately 100× less trainable-parameter storage. Figure 6 shows accuracy vs. shot curves; Figure 7 plots the complexity-accuracy trade-off.

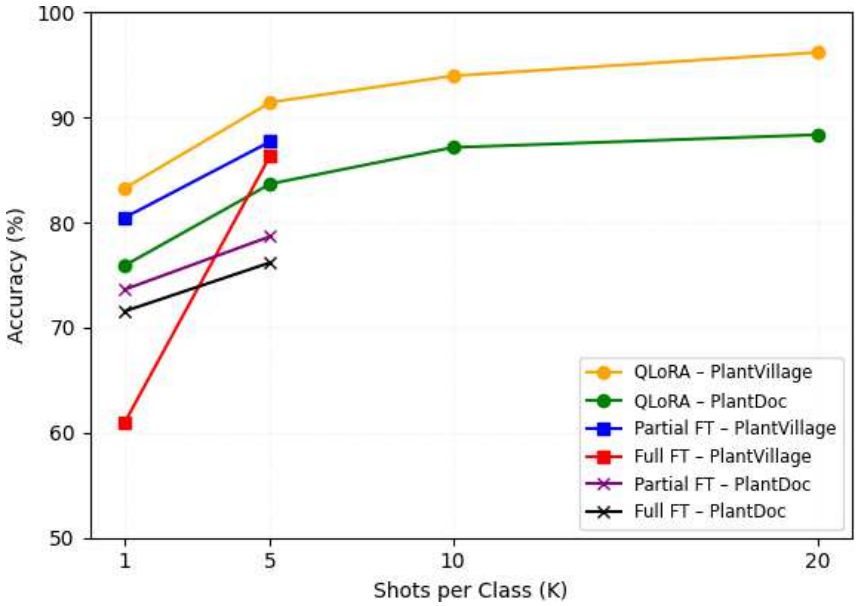


Figure 6. Accuracy versus number of shots for QLoRA, Partial Fine-Tune, and Full Fine-Tune on PlantVillage and PlantDoc datasets.

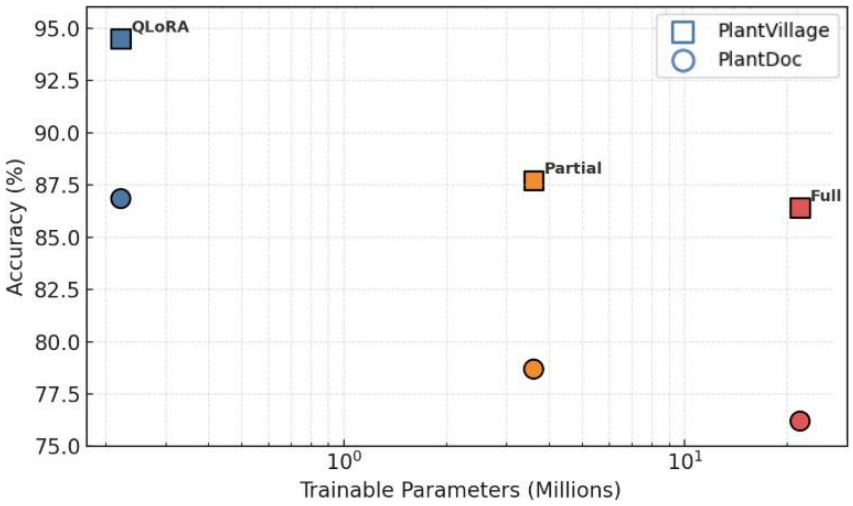


Figure 7. Trade-off between model complexity and accuracy in the 5-way 5-shot setting. QLoRA achieves the highest accuracy with only ~0.22 M parameters, whereas Partial and Full FT require 3.6 M and 21.8 M parameters respectively, sacrificing performance on PlantDoc as complexity grows.

5.3 Feature Space Visualization

To provide qualitative insight into the discriminative capability of the framework, we visualized the feature embeddings using t-Distributed Stochastic Neighbor Embedding (t-SNE). Figure 8 illustrates the feature distribution for a randomly sampled 5-way 5-shot episode from the PlantVillage test set.

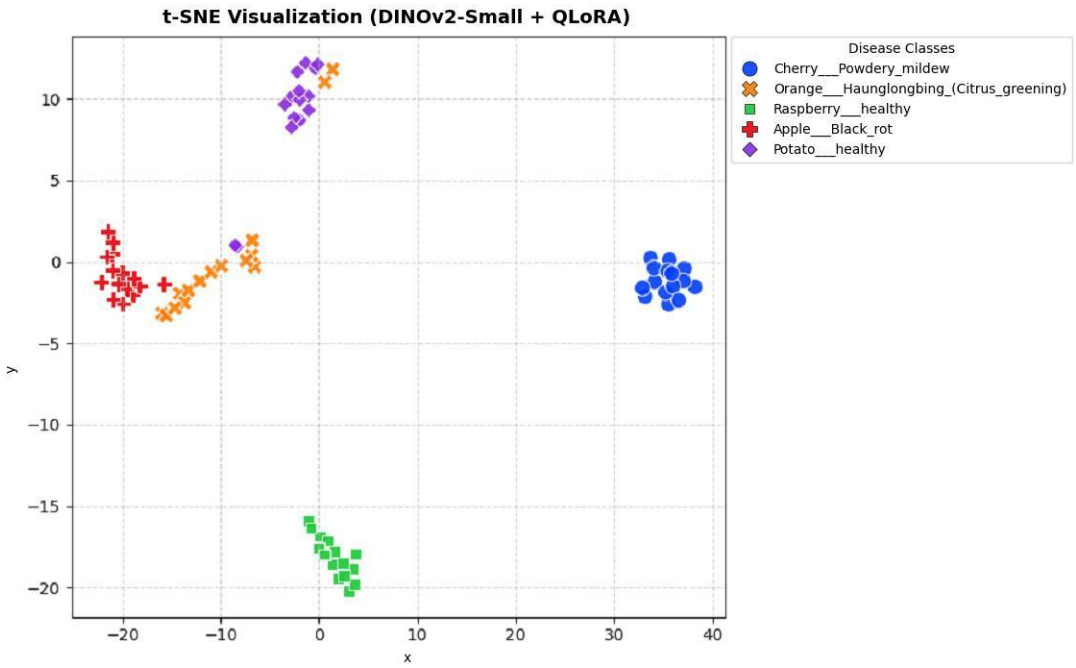


Figure 8. t-SNE visualization of the feature embeddings generated by the proposed DINOv2-S + QLoRA model for a randomly sampled 5-way 5-shot task. The distinct color clusters correspond to real disease classes from the PlantVillage dataset. The clear separation and compactness of these clusters visually confirm the model's ability to learn robust and discriminative representations for few-shot classification.

The DINOv2-S + QLoRA model produces compact and well-separated clusters for distinct disease categories. The clear boundaries between clusters indicate high inter-class variance and low intra-class variance, which is favorable for prototype-based classification. This visual evidence is consistent with the quantitative accuracy reported in Table 2.

5.4 Comparison with Prior Work

Table 4 benchmarks our method against prior studies.

Table 4. Comparison with prior methods on PlantVillage and PlantDoc datasets (5-way classification).

Study	Backbone	Img	PV 1s	PV 5s	PD 1s	PD 5s
Rezaei et al. (2024)	ViT-Base	224	–	94.09	–	86.79
Ullah et al. (2022)	CNN (Various)	128	64.10	80.90	24.10	26.30
Chen et al. (2021)	LFM-CNAPS	84	–	82.40	–	–
Mu et al. (2024)	Contrastive+Proto	224	–	–	–	79.50*
Wu et al. (2023)	Meta-baseline	84	72.30	86.20	–	–
ProtoNet baseline	ResNet-50	224	83.24	92.15	76.79	84.12

ProtoNet baseline	ResNet-50	128	81.05	89.45	73.15	81.77
Ours	DINOv2-S+QLoRA	224	85.46	94.51	79.05	86.85
Ours	DINOv2-S+QLoRA	128	83.27	91.45	75.91	83.69

* Mu et al. (2024) report accuracy on a different real-world plant disease dataset; the value is included for reference but is not directly comparable due to different class compositions.

At 224×224 resolution in the 5-way 5-shot setting, our method achieves 94.51 ± 0.35% on PlantVillage and 86.85 ± 0.42% on PlantDoc. The PlantVillage result represents a 0.42 percentage point improvement over Rezaei et al. (2024). On PlantDoc, the difference is 0.06 percentage points, which falls within the confidence interval range and should be interpreted as comparable performance rather than a definitive improvement. Importantly, our approach achieves these results using approximately 221K trainable parameters compared to 86M in Rezaei et al. (2024), representing a ~390× reduction, and 10,000 training episodes compared to 100,000, a 10× reduction in training iterations. The ResNet-50 ProtoNet baseline at 224×224 achieves 92.15% on PlantVillage and 84.12% on PlantDoc, indicating that much of the performance gain derives from the DINOv2 backbone's representation quality; While the DINOv2 backbone provides strong representations, a direct comparison with a frozen DINOv2 baseline would further isolate the exact contribution of QLoRA adaptation and remains a direction for future work. We note that direct comparison across studies is complicated by differences in class splits, episode construction, evaluation protocols, and backbone model sizes. Where available, we have aligned our experimental setup with the protocols reported in prior work, particularly that of Rezaei et al. (2024). Ullah et al. (2022) results are drawn from the Meta-Album benchmark under the matching N-way K-shot protocol; Chen et al. (2021) results correspond to LFM-CNAPS evaluated on PlantVillage with 84×84 resolution; Wu et al. (2023) results are from their meta-baseline evaluation on the PlantVillage subset; and Mu et al. (2024) report on a separate real-world dataset and are included for contextual reference only (marked with * in Table 4). Studies for which experimental settings differ substantially are noted in the table footnote.

5.5 Ablation Studies

5.5.1 Effect of Number of Shots

As K increases from 1 to 20, accuracy improves across both datasets (Table 2). The most substantial gains occur between 1-shot and 5-shot settings, with PlantVillage accuracy rising from 83.27% to 91.45% and PlantDoc from 75.91% to 83.69%. Returns diminish at higher shot counts, with the 10-to-20 shot improvement being approximately 2 percentage points on each dataset.

5.5.2 Effect of Backbone Architecture

To assess the impact of model capacity, we compared DINOv2-Small against DINOv2-Base (approximately 86M parameters). Table 5 presents the results.

Table 5. Backbone capacity comparison (5-way 5-shot, 128×128).

Backbone	**Params**	**Trainable**	**Trainable-Param. Storage (MB)**	**PV Acc (%)**	**PD Acc (%)**
DINOv2-Small	~21M	221K	~2.53	91.45±0.38	83.69±0.49
DINOv2-Base	~86M	442K	~5.06	94.21±0.33	87.42±0.43

Scaling to DINOv2-Base improves accuracy by 2.76 and 3.73 percentage points on PlantVillage and PlantDoc, respectively, at the cost of doubling the trainable parameter count and trainable-parameter storage. Since DINOv2-Small already achieves competitive results with substantially lower resource requirements, we retain it as the default backbone, noting that DINOv2-Base may be preferable when computational constraints are less restrictive.

5.5.3 Effect of Input Image Resolution

Table 6 presents results at different input resolutions.

Table 6. Effect of input resolution (DINOv2-S + QLoRA + ProtoNet, 5-way).

Setting	Image Size	PV Acc (%)	PD Acc (%)	
5w 1s	128×128	83.27±0.52	75.91±0.61	
5w 5s	128×128	91.45±0.38	83.69±0.49	
5w 1s	518×518	89.44±0.44	79.34±0.56	
5w 5s	518×518	95.92±0.27	87.98±0.39	
5w 5s	224×224	94.51±0.35	86.85±0.42	

Higher resolutions consistently improve accuracy, with 518×518 yielding the best results. The improvement from 128×128 to 224×224 is 3.06 points on PlantVillage and 3.16 on PlantDoc for 5-shot, indicating that finer spatial detail benefits disease feature extraction. The 224×224 resolution offers a practical compromise between accuracy and computational cost.

5.5.4 Effect of LoRA Rank

To understand the sensitivity to the rank parameter r, we conduct experiments with $r \in \{2, 4, 8, 16, 32\}$ while keeping $\alpha = 2r$ (following the convention of scaling α proportionally to r). Table 7 presents the results.

Table 7. Effect of LoRA rank on accuracy and parameter count (5-way 5-shot, 128×128).

Rank (r)	α	Trainable Params	Trainable-Param. Storage (MB)	PV Acc (%)	PD Acc (%)
2	4	~56K	~0.64	89.82±0.42	81.45±0.53
4	8	~111K	~1.27	90.71±0.40	82.88±0.51
8	16	~221K	~2.53	91.45±0.38	83.69±0.49
16	32	~442K	~5.06	91.62±0.37	83.78±0.48
32	64	~884K	~10.12	91.55±0.38	83.71±0.49

Accuracy increases substantially from r = 2 to r = 8, with diminishing returns beyond r = 8. Ranks of 16 and 32 yield marginal improvements of 0.17% and 0.10% on PlantVillage relative to r = 8, while doubling and quadrupling the parameter count, respectively. This confirms r = 8 as an effective default that balances adaptation capacity with parameter efficiency.

5.5.5 Quantization Effect: QLoRA vs. Standard LoRA

To isolate the contribution of 4-bit quantization, Table 3 includes a direct comparison between QLoRA (4-bit backbone) and standard LoRA (FP16 backbone) using identical adapter configurations ($r = 8$, $\alpha = 16$). Across all shot settings and both datasets, QLoRA and LoRA achieve similar accuracy, with QLoRA showing a consistent but small advantage of 0.2–0.5 percentage points. This suggests that 4-bit quantization does not degrade classification performance for this task, and may provide a mild regularization effect. The primary benefit of quantization is the reduction in backbone storage requirements, from approximately 42 MB (FP16) to approximately 10.5 MB (NF4), facilitating deployment on storage-constrained hardware.

6. Discussion

The experimental results validate the effectiveness of integrating QLoRA with DINOv2 and Prototypical Networks for few-shot plant disease classification. In this section, we discuss the key findings, their implications, and limitations of the current work.

Performance Analysis. Quantitative analysis (Table 2) reveals that the model achieves high accuracy and balanced F1-scores across both 3-way and 5-way settings. The close alignment between Precision and Recall values (e.g., 96.22% Precision vs. 96.28% Recall in the 5-way 20-shot setting on PlantVillage) indicates that the model does not exhibit systematic class bias. The t-SNE visualization (Figure 8) provides qualitative support for this finding, showing distinct and compact class clusters in the learned embedding space. The clear separation in feature space is consistent with the metric-learning objective and suggests that QLoRA adaptation preserves the discriminative characteristics of the pre-trained backbone.

Role of DINOv2 vs. QLoRA Adaptation. An important consideration is the relative contribution of the DINOv2 backbone versus the QLoRA adaptation. The ResNet-50 ProtoNet baseline at 224×224 achieves 92.15% on PlantVillage (Table 4), indicating that backbone quality is a substantial factor. The DINOv2-S + QLoRA framework achieves 94.51% under the same conditions, representing a 2.36 percentage point improvement. This suggests that while DINOv2's self-supervised features provide a strong foundation, the QLoRA adaptation contributes meaningful additional task-specific refinement. The comparison with frozen DINOv2 features (without any adaptation) would further isolate the QLoRA contribution and represents a direction for future investigation.

Comparison with State-of-the-Art. On PlantVillage, our method achieves a 0.42 percentage point improvement over Rezaei et al. (2024) in the 5-way 5-shot setting. On PlantDoc, the improvement is 0.06 percentage points, which is within the confidence interval and should be interpreted as comparable performance. The meaningful comparison between these two approaches lies in their efficiency profiles: our method achieves similar accuracy with approximately 390× fewer trainable parameters and 10× fewer training episodes. This efficiency advantage is the primary practical contribution, particularly for scenarios where computational resources are limited.

Why QLoRA Outperforms Full Fine-Tuning. The consistent advantage of QLoRA over Full Fine-Tuning across all shot settings (Table 3) can be attributed to the regularization effect of constraining the parameter space. In few-shot learning, the limited number of training examples per episode makes models with large parameter counts susceptible to overfitting. Full Fine-Tuning updates all 21.8M parameters based on only K examples per class, resulting in poor generalization, especially in 1-shot settings where it achieves only 60.96% on PlantVillage. By restricting updates to 221K parameters through low-rank decomposition, QLoRA preserves the generalizable features learned during pre-training while allowing sufficient adaptation for task-specific discrimination. This interpretation is supported by the observation that Partial Fine-Tuning (3.6M parameters) performs between Full Fine-Tuning and QLoRA, consistent with the hypothesis that constraining the update space improves generalization in low-data regimes.

Efficiency Considerations. The QLoRA framework requires an approximate trainable-parameter storage of 2.53 MB, compared to approximately 42 MB for Partial Fine-Tuning and 250 MB for Full Fine-Tuning (Table 1). Peak GPU memory during training was measured at approximately 4.1 GB for QLoRA, 8.2 GB for Partial Fine-Tuning, and 14.6 GB for Full Fine-Tuning on A100 hardware. Inference on NVIDIA A100 hardware takes approximately 0.20 ms per

image at 224×224 resolution. These measurements are specific to A100 GPUs; actual performance on resource-constrained devices such as embedded processors, agricultural drones, or mobile phones would depend on model quantization compatibility, hardware-specific inference engines, and memory bandwidth. We note that the 4-bit quantized backbone is compatible with emerging edge inference frameworks (e.g., ONNX Runtime, TensorRT with INT4 support), suggesting potential for edge deployment, though this requires empirical validation in future work.

Limitations. Several limitations of this study should be noted. First, our comparisons with alternative PEFT methods are limited to standard LoRA, Partial Fine-Tuning, and Full Fine-Tuning. Direct comparison with visual prompt tuning (VPT), adapter modules, and BitFit would provide a more comprehensive picture of QLoRA's relative merits and is planned for future work. Second, the evaluation is conducted exclusively on plant disease datasets. While this focus is appropriate for the application domain, generalization of our findings to other few-shot vision tasks remains to be demonstrated. Third, the current episodic training assumes a fixed set of classes during meta-testing; continuous learning scenarios where new diseases appear sequentially were not modeled. Fourth, although we report inference times on A100 GPUs, actual edge device benchmarking has not been performed. Finally, per-class analysis, including confusion matrices that would reveal which specific disease pairs are most challenging, was not included in the current evaluation and represents a valuable addition for subsequent work.

7. Conclusion

In this paper, we presented a few-shot learning framework for plant disease classification that combines a DINOv2-S Vision Transformer backbone, QLoRA-based parameter-efficient adaptation, and Prototypical Network classification. The framework fine-tunes approximately 1% of model parameters (~221K) with an approximate trainable-parameter storage of 2.53 MB, representing a reduction of approximately 100× relative to full fine-tuning.

Experiments on the PlantVillage and PlantDoc datasets demonstrate competitive performance in the 5-way 5-shot setting, with accuracies of 94.51 ± 0.35% and 86.85 ± 0.42% at 224×224 resolution, respectively. Ablation studies confirm the robustness of the approach across varying shot numbers, image resolutions, backbone architectures, and LoRA rank configurations. Comparison with standard LoRA (without quantization) indicates that 4-bit quantization does not degrade accuracy while providing substantial storage savings.

These results suggest that parameter-efficient adaptation of vision transformers offers a practical approach for few-shot plant disease classification in data-scarce agricultural settings. Future work will focus on direct comparison with additional PEFT methods (VPT, adapter modules), empirical evaluation on edge hardware platforms, integration of continual learning strategies for sequential disease class addition, and per-class error analysis to identify challenging disease pairs and inform targeted improvements.

Ethics Statement

All datasets used are publicly available and fully de-identified. No additional human or animal data were collected. This work complies with all relevant ethical guidelines and the data providers' terms of use. No conflicts of interest are declared.

Funding Declaration

The authors received no financial support for the research, authorship, and/or publication of this study.

Data Availability Statement

The PlantVillage dataset is available through TensorFlow Datasets. The PlantDoc classification dataset can be obtained from https://github.com/pratikkayal/PlantDoc-Dataset. To facilitate reproducibility, the exact class splits,

random seeds used across all three evaluation runs, episodic sampling configuration, and implementation details (including bitsandbytes and PEFT library versions) will be released as supplementary materials upon acceptance.

CRediT Author Statement

All authors contributed equally to conceptualization, methodology, investigation, writing, and revision of this manuscript.

References

Bayat Toksöz, S., & Işık, G. (2025). Efficient Adaptation of Large Language Models for Sentiment Analysis: A Fine-Tuning Approach. Journal of the Institute of Science and Technology, 15(4), 1149–1164.

Bayat Toksöz, S., & Işık, G. (2026). Parameter-efficient fine-tuning of LLaMA models for financial sentiment classification. Cluster Computing, 29(1), 41. [In press]

Chen, L., Cui, X., & Li, W. (2021). Meta-learning for few-shot plant disease detection. Foods, 10(10), 2441.

de Andrade Porto, J. V., et al. (2023). Usage of few-shot learning and meta-learning in agriculture: A literature review. Smart Agricultural Technology, 5, 100307.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36, 10088–10115.

Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021.

Fan, Q., Tang, C.-K., & Tai, Y.-W. (2022). Few-shot object detection with model calibration. ECCV, 720–739.

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318.

Günde, M., & Işık, G. (2025). A Robust Apple Disease Classification Approach Using Vision Transformer-Based Meta-Learning. Journal of the Institute of Science and Technology, 15(4), 1193–1205.

Günde, M., Keskin, F., & Işık, G. (2026). Empowering few-shot plant disease diagnosis with DeiT and MAML++. Cluster Computing, 29(1), 13. [In press]

Gündüz, M. Ş., & Işık, G. (2023). A new YOLO-based method for social distancing from real-time videos. Neural Computing and Applications, 35(21), 15261–15271.

Hiller, M., Ma, R., Harandi, M., & Drummond, T. (2022). Rethinking generalization in few-shot classification. NeurIPS, 35, 3582–3595.

Houlsby, N., et al. (2019). Parameter-efficient transfer learning for NLP. ICML, 2790–2799.

Hu, E. J., et al. (2022). LoRA: Low-rank adaptation of large language models. ICLR.

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

Işık, G. (2023). Improving plant disease recognition through gradient-based few-shot learning with attention mechanisms. Journal of the Institute of Science and Technology, 13(3), 1482–1495.

Işık, G., & Paçal, I. (2024). Few-shot classification of ultrasound breast cancer images using meta-learning algorithms. Neural Computing and Applications, 36(20), 12047–12059.

Jia, M., et al. (2022). Visual prompt tuning. ECCV, 709–727.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980.

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. ICML Deep Learning Workshop.

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419.

Mu, J., et al. (2024). Few-shot disease recognition algorithm based on supervised contrastive learning. Frontiers in Plant Science, 15, 1341831.

Oquab, M., et al. (2023). DINOv2: Learning robust visual features without supervision. arXiv:2304.07193.

Rezaei, M., et al. (2024). Plant disease recognition in a low data scenario using few-shot learning. Computers and Electronics in Agriculture, 219, 108812.

Richard, B., Qi, A., & Fitt, B. D. L. (2022). Control of crop diseases through Integrated Crop Management. Plant Pathology, 71(1), 187–206.

Singh, D., et al. (2020). PlantDoc: A dataset for visual plant disease detection. ACM IKDD CoDS and COMAD, 249–253.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. NeurIPS, 30.

Su, M., He, F., Li, G., & Li, F. (2025). PrototypeFormer: Learning to explore prototype relationships for few-shot image classification. Neurocomputing, 130326.

Sun, J., et al. (2024). Few-shot learning for plant disease recognition: A review. Agronomy Journal, 116(3), 1204–1216.

Ullah, I., et al. (2022). Meta-album: Multi-domain meta-dataset for few-shot image classification. NeurIPS, 35, 3232–3247.

Uzhinskiy, A. (2025). Evaluation of Different Few-Shot Learning Methods in the Plant Disease Classification Domain. Biology, 14(1), 99.

Vinyals, O., et al. (2016). Matching networks for one shot learning. NeurIPS, 29.

Wu, X., et al. (2023). Meta-learning shows great potential in plant disease recognition under few available samples. The Plant Journal, 114(4), 767–782.

Yang, S., et al. (2024). From laboratory to field: Cross-domain few-shot learning for crop disease identification. Frontiers in Plant Science, 15, 1434222.

Zaken, E. B., Goldberg, Y., & Ravfogel, S. (2022). BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. ACL, 1–9.