

Evaluating Few-Shot Learning for Cherry Leaf Disease Classification in Real-World Agricultural Settings

Dostonbek Nazirzhonov¹[0009–0008–6845–4056], Md Easin Arafat²[0000–0003–4014–9144], Nazrul Islam³[0000–0002–9276–8388], and Utkarsh Mall⁴[0009–0002–5302–0657]

¹ Faculty of Informatics,
Eötvös Loránd University, Pázmány Péter str. 1/C, Budapest-1117, Hungary
`c0cshf@inf.elte.hu`

² Department of Data Science and Engineering, Faculty of Informatics,
Eötvös Loránd University, Pázmány Péter str. 1/C, Budapest-1117, Hungary
`arafatmdeasin@inf.elte.hu`

³ Department of Information and Communication Technology,
Mawlana Bhashani Science and Technology University, Tangail, Bangladesh
`nazrul@mbstu.ac.bd`

⁴ Department of Computer Vision,
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
`utkarsh.mall@mbzuai.ac.ae`

Abstract. We address the challenge of limited data and domain shift in plant disease recognition by introducing a 400-image cherry leaf dataset spanning four disease categories, collected on real farms in Kyrgyzstan. Images exhibit natural lighting variation, background clutter, and Uzbek-language class labels, providing a realistic test bed for few-shot and zero-shot methods. We evaluate fine-tuning, Prototypical Networks, k-NN, and zero-shot CLIP using ResNet50 and ViT-B/16 backbones in 4-way 1-shot and 5-shot settings. Fine-tuning ResNet50 achieves the best performance (81.32% in 5-shot), while Prototypical Networks and k-NN offer stable, competitive results. ViT-B/16 and CLIP degrade under low-shot conditions and non-English labels; zero-shot CLIP drops from 40.50% with English prompts to 28.75% with Uzbek prompts, quantifying language bias in a real agricultural deployment scenario. These results establish practical baselines for robust recognition in real-world, data-scarce agricultural settings.

Keywords: Few-shot Learning · Plant Disease Classification · Computer Vision in Agriculture · Cross-Lingual Learning · Vision Transformer.

1 Introduction

Advances in Computer Vision (CV) have expanded machine capability to interpret visual information from images and videos [1], powering applications

like face detection, medical diagnosis, and plant disease recognition. Training effective CV models typically requires large labeled datasets [2], which are often expensive or infeasible to collect in agricultural settings. Few-shot learning (FSL) addresses this by enabling models to learn new classes from only a few examples [3, 4]. However, few-shot methods often overfit to curated benchmarks and generalize poorly to real-world agricultural tasks [3, 16].

We introduce a dataset of 400 cherry leaf images, evenly split across four disease categories (Sarik, Teshik, Chirish, Soglom) collected outdoors on farms in Kyrgyzstan. Unlike lab-captured datasets [5, 6], ours includes natural shadows, occlusion, uneven lighting, cluttered backgrounds, and Uzbek class labels, making it a strong test bed for non-Western, real-world deployment. Using this dataset, we systematically evaluate fine-tuning, Prototypical Networks [7], and k-NN [8] with ResNet50 [9] and ViT-B/16 [10] backbones, and assess zero-shot CLIP [11] to examine generalization and language bias from non-English labels.

Our contribution is a rigorous empirical evaluation of established few-shot methods on messy, real-world data, providing practical baselines for generalization, robustness, and cross-lingual bias research. Specifically, we contribute: a real-world cherry leaf disease dataset of 400 field images with Uzbek-language labels capturing natural variation absent from curated benchmarks; a systematic comparison of three few-shot methods across two backbones under 1-shot and 5-shot settings; a cross-lingual bias analysis for zero-shot CLIP contrasting English and Uzbek prompts; and practical deployment guidelines for resource-constrained agricultural settings.

The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 describes the dataset; Section 4 details the models, methods, and protocol; Section 5 reports results; Section 6 discusses implications; and Section 7 presents conclusions and limitations.

2 Related Work

2.1 Few-Shot Learning and Metric-Based Methods

Few-shot learning is important where large annotated datasets are unavailable [3, 4]. Solutions fall into three families: metric-based models, optimization-driven meta-learning, and fine-tuning-based transfer learning.

Prototypical Networks [7] compute a class prototype and classify queries by Euclidean distance; we evaluate this method with ResNet50 [9] and ViT-B/16 [10] backbones. Matching Networks [12] and Relation Networks [13] require episodic meta-training across diverse tasks, which is infeasible for our single-domain dataset of 400 images, so we exclude them and focus on frozen pretrained features. k-Nearest Neighbors (k-NN) classifiers have also regained attention in few-shot settings [8]; we include k-NN with ResNet50 and ViT-B/16 embeddings as a baseline.

2.2 Fine-Tuning and Transfer Learning

Optimization-based meta-learning (e.g., MAML [14]) is computationally intensive and struggles under large task-distribution shifts, so we exclude it for the same reason. Fine-tuning pretrained ImageNet models [2] remains a practical alternative that can match sophisticated meta-learning under moderate domain shift, and is central to our experiments, adapting ImageNet-pretrained backbones using the few support samples available from our cherry leaf dataset.

2.3 Zero-Shot Learning and CLIP

Zero-shot learning recognizes classes with no labeled training samples, typically via large vision-language models. CLIP [11] classifies images by matching them to the closest text prompt without task-specific fine-tuning. However, CLIP’s training data is heavily English-centered, introducing linguistic bias for low-resource languages [11, 19]. Multilingual variants such as Multilingual-CLIP [19], AltCLIP [20], and NLLB-CLIP [24] still underperform relative to English due to limited low-resource training data. We retain the original CLIP as our baseline to measure language bias in the most widely deployed vision-language model.

2.4 Plant Disease Classification with Deep Learning

Annotated plant disease datasets are scarce and costly, making few-shot learning increasingly relevant in agriculture. Prior work shows leaf-image approaches are often evaluated on inconsistent datasets [15], laboratory-trained models fail to transfer to field conditions [16], and uncontrolled UAV-based capture adds further difficulty [17], motivating evaluation of few-shot methods under realistic agricultural conditions.

3 Dataset Description

We created a dataset of cherry tree leaf images collected on outdoor farms in Kyrgyzstan using an iPhone 14 over a single day in July, across three farms in the village of Uch-Korgon (Kadamzhay District, Batken Region). All images were taken under working-farm conditions, naturally including variations in illumination, leaf orientation, shadows, occlusions, and background clutter. Data collection occurred during a single day (morning until afternoon); each photograph captures a distinct leaf, and individual leaves were sampled only once. The final dataset contains 400 images distributed evenly across four categories, summarized in Table 1; representative images from each class are shown in Fig. 1.

Although 400 images is small by the standards of fully supervised training, it is appropriate for the episodic few-shot protocol used here: a model adapts to only one or five labeled images per class, and the remaining images serve as a query pool for evaluation. Each class includes exactly 100 images with labels written in Uzbek, reflecting local grower terminology from Uch-Korgon rather

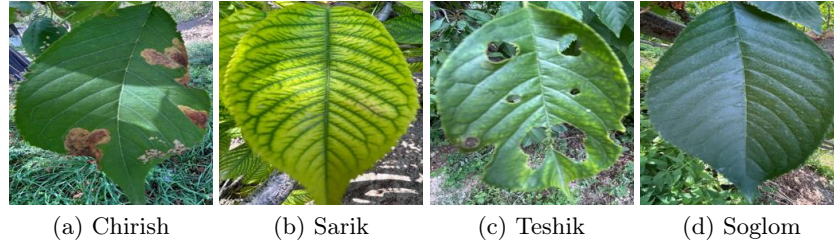


Fig. 1. Representative images from the four cherry leaf disease classes: (a) Chirish (brown rot disease with leaf decay), (b) Sarik (leaf yellowing), (c) Teshik (shot hole disease with perforations), and (d) Soglom (healthy leaf).

Table 1. The four cherry leaf classes. Class names are given in Uzbek, the language used on the collecting farms.

Uzbek	English gloss	#	Img. Visual description
Sarik	“yellow”	100	Diffuse leaf yellowing (chlorosis)
Teshik	“holed”	100	Perforations from shot-hole disease
Chirish	“rot”	100	Brown rot with decaying tissue
Soglom	“healthy”	100	Uniform green, disease-free leaf

than diagnoses independently verified by a certified agronomist; we treat this as a limitation (Section 7). Source photographs were captured at 960×1280 pixels, manually annotated using makesense.ai [18], cropped to the central leaf, and resized to 224×224 pixels for all models. The full dataset is publicly available on Mendeley Data ⁵.

4 Methodology

4.1 Overview

We evaluate three few-shot methods (Prototypical Networks [7], fine-tuning, and k-NN [8]) combined with two backbones (ResNet50 [9] and ViT-B/16 [10]) on 4-way 1-shot and 5-shot cherry leaf disease classification, and additionally examine zero-shot CLIP [11] with Uzbek class labels. The overall pipeline, shown in Fig. 2, proceeds from field images through preprocessing, feature extraction, classification, and evaluation.

4.2 Data Preprocessing

Each image was cropped to the central leaf using human-annotated bounding boxes, resized to 224×224 pixels, and normalized using ImageNet mean and standard deviation [2]. No data augmentation was applied, so images retain the natural variability of field capture; augmentation is left for future work (Section 7).

⁵ <https://data.mendeley.com/datasets/wp3b6pz9gc/3>

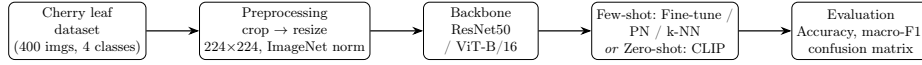


Fig. 2. Overview of the experimental pipeline: field images are cropped and normalized, encoded by a ResNet50 or ViT-B/16 backbone, classified either by a few-shot method (fine-tuning, Prototypical Networks, or k-NN) or by zero-shot CLIP, and evaluated with accuracy, macro-averaged F1, and confusion matrices.

4.3 Backbone Models

ResNet50 ResNet50 [9] is a 50-layer convolutional network whose central innovation is the residual (skip) connection: instead of learning a full transformation, each block learns a residual $F(x)$, so the block output is

$$y = \text{ReLU}(F(x) + x), \quad (1)$$

where the identity term x stabilizes gradient flow during backpropagation. We use ResNet50 pretrained on ImageNet-1K [2] and extract a 2048-dimensional global-average-pooled (GAP) embedding. As shown in Fig. 3, the stem, Layer1, and Layer2 remain frozen; Layer3, Layer4, and the fully connected head are fine-tuned when the fine-tuning baseline is used, while the GAP embedding is passed directly to Prototypical Networks or k-NN for metric-based evaluation.

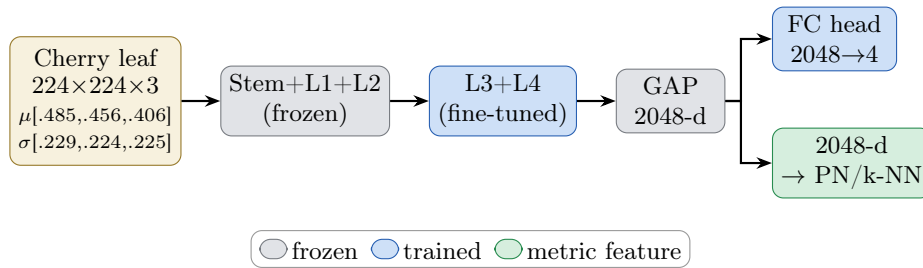


Fig. 3. ResNet50 usage. Stem, Layer1, and Layer2 are frozen; Layer3, Layer4, and the FC head are fine-tuned. The 2048-d GAP embedding feeds either the FC head (fine-tuning) or the metric methods (PN, k-NN) directly.

ViT-B/16 ViT-B/16 [10] splits an image into 16×16 patches, linearly embeds each patch into a token, and processes the resulting sequence with 12 stacked transformer blocks of multi-head self-attention and feed-forward layers, producing a 768-dimensional class-token (CLS) embedding from approximately 86M parameters. As shown in Fig. 4, the entire transformer backbone is kept frozen in our experiments; for fine-tuning we train only a linear-probe head mapping the CLS token to four classes, while Prototypical Networks and k-NN use the CLS embedding directly.

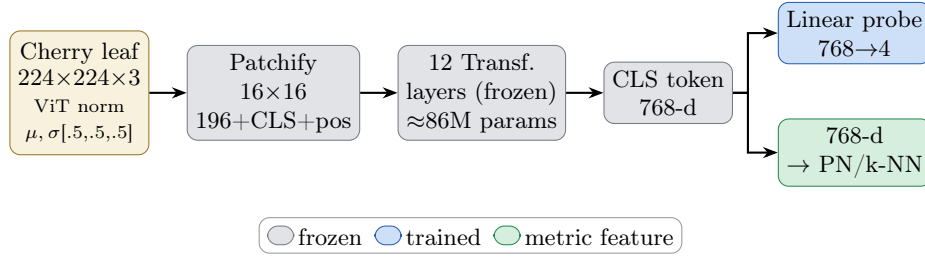


Fig. 4. ViT-B/16 usage. The full transformer backbone is frozen; only the linear-probe head is trained for fine-tuning, while PN and k-NN use the CLS embedding directly.

CLIP CLIP [11] comprises a ViT-B/32 image encoder and a Text Transformer, trained jointly via a contrastive objective that aligns image-text pairs in a shared embedding space. In zero-shot classification, each class name is encoded as a text prompt to form a class prototype, and the query image embedding is compared to all prototypes by cosine similarity. Fig. 5 illustrates both encoders used off the shelf with no additional training.

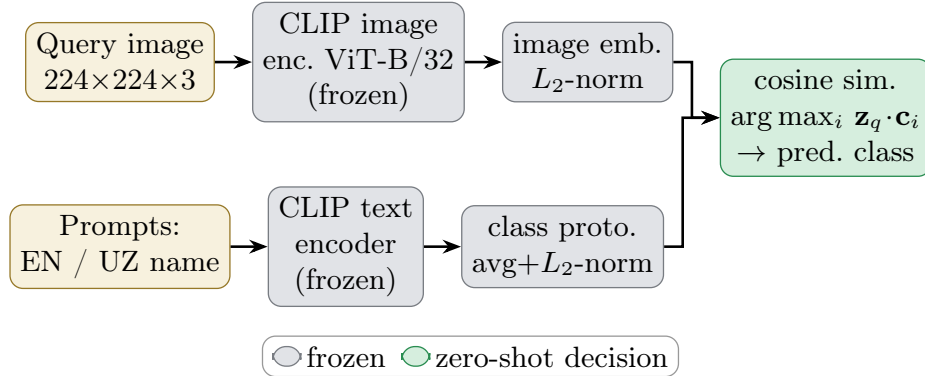


Fig. 5. CLIP zero-shot classification. Both encoders are frozen. Text prompts are encoded, L_2 -normalized, and averaged into a class prototype; the predicted label maximises cosine similarity to the query-image embedding.

4.4 Few-Shot Learning Methods

Prototypical Networks Prototypical Networks [7] classify images by comparing them to class prototype vectors computed as the mean embedding of support samples. For class k with support set S_k , the prototype is

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_{\theta}(x_i), \quad (2)$$

where f_θ is the feature extractor. Each query is then assigned to the class whose prototype minimizes the squared Euclidean distance,

$$\text{Distance}(\mathbf{q}, \mathbf{c}_i) = \|\mathbf{q} - \mathbf{c}_i\|_2^2, \quad (3)$$

with backbone weights held frozen throughout, isolating the quality of the pre-trained representations. ResNet50 provides 2048-dimensional GAP features; ViT-B/16 provides the 768-dimensional CLS token (see Figs. 3 and 4).

Fine-Tuning Within each episode, the cross-entropy loss

$$\mathcal{L}_{\text{CE}} = - \sum_{i=1}^C y_i \log(p_i) \quad (4)$$

is minimized on the support images only, with evaluation always performed on disjoint query samples. For ResNet50 we unfreeze Layer3, Layer4, and the FC head; for ViT-B/16 we apply strict linear probing, training only the classification head. These strategies are not identical, so part of the observed fine-tuning gap between backbones may stem from the differing adaptation strategy rather than the architecture alone; we therefore treat the fine-tuning comparison as indicative rather than strictly controlled, and return to this confound in Section 7. Full hyperparameter and k-NN configuration details are given in Table 2; the learning rate was selected via minimal preliminary experiments, and all other choices follow established practice [10, 21].

Table 2. Fine-tuning hyperparameters (top) and k-NN configuration (bottom), reported as 1-shot/5-shot where they differ. LR: learning rate; SupBatch/QryBatch: support/query batch size; L3/L4/FC: residual blocks 3 and 4 and the FC head. ViT-B/16 optimizer settings were not retained after the experiments (n.r.); authors should replace n.r. entries with the actual values before final submission to support reproducibility.

<i>Fine-tuning hyperparameters</i>		
Parameter	ResNet50 (1/5)	ViT-B/16 (1/5)
Trainable	L3, L4, FC	Head only
Optimizer	SGD (mom. 0.9)	n.r.
LR	0.01	n.r.
Weight decay	5e-4	n.r.
Epochs	50	n.r.
SupBatch	4 / 20	4 / 20
QryBatch	Full	32
<i>k-NN configuration</i>		
Experiment	1-shot	5-shot
k	1	5
Weighting	Distance	
Support total	4	20

k-Nearest Neighbors k-NN [8] classifies each query by voting among the k nearest support samples in the embedding space. Given query embedding $\mathbf{x}_q$ and support set S , the Euclidean distance is

$$d(\mathbf{x}_q, \mathbf{x}_i) = \|f(\mathbf{x}_q) - f(\mathbf{x}_i)\|_2, \quad (5)$$

and the predicted label is

$$\hat{y}_q = \arg \max_c \sum_{i \in N_k(x_q)} \mathbf{1}[y_i = c], \quad (6)$$

where $N_k(x_q)$ denotes the k nearest neighbors. All embeddings are L_2 -normalized before distance computation, and distance-weighted voting (`weights='distance'`) is used throughout. In the 1-shot case ($k=1$), k-NN and Prototypical Networks are mathematically equivalent since a single support example is its own prototype. The full configuration is given in Table 2 (bottom).

4.5 Experimental Protocol

Few-Shot Protocol All six method-backbone combinations were evaluated under a unified episodic protocol using 4-way 1-shot and 5-shot settings, executed in Google Colaboratory with a T4 GPU. In 1-shot, the support set contains 4 images (one per class) and the query set contains the remaining 396; in 5-shot, the support set contains 20 images, and the query set contains 380. Results were averaged over 10 episodes with a fixed seed of 42. Although 10 episodes are below the 600-1000 used on standard benchmarks [7], evaluating each episode on the entire remaining dataset rather than a fixed query set of 15 per class removes most query-sampling variance [23], so each episode already yields a low-variance accuracy estimate. We note that only the support set is re-sampled across episodes; the query set consists of all remaining images and therefore overlaps heavily between episodes, so the 10 episode-level estimates are not statistically independent, and the reported 95% CIs, which primarily reflect variance in support-set composition, likely understate total uncertainty.

Zero-Shot Protocol CLIP (ViT-B/32) was used entirely off the shelf with no fine-tuning [11]. For each class we prepared two text prompts (Table 3): the English translation and the Uzbek class name, each inserted into the template “a photo of a {prompt}”. We fixed a single template across languages rather than an ensemble of templates or prompt tuning, so that the English-Uzbek comparison isolates language bias rather than confounding it with prompt-selection effects; we note this as a limitation, since CLIP accuracy is known to be sensitive to prompt phrasing, and leave a full prompt-sensitivity study to future work. Text embeddings were L_2 -normalized and averaged per class to form a class prototype,

$$\mathbf{c}_i = \frac{1}{N} \sum_{j=1}^N \frac{f_{\text{text}}(t_j)}{\|f_{\text{text}}(t_j)\|_2}, \quad i = 1, \dots, 4. \quad (7)$$

Query images were encoded, L_2 -normalized,

$$\mathbf{z}_q = \frac{f_{\text{vision}}(x_q)}{\|f_{\text{vision}}(x_q)\|_2}, \quad (8)$$

and assigned to the class with the highest cosine similarity,

$$\hat{y}_q = \arg \max_i \mathbf{z}_q \cdot \mathbf{c}_i. \quad (9)$$

We report three conditions: English prompts, Uzbek prompts, and a multilingual ensemble that averages English and Uzbek embeddings per class before re-normalization. All 400 images were evaluated in a single inference pass.

Table 3. Representative CLIP text prompts per class (each inserted into “a photo of a {prompt}”).

Uzbek name	English translation
Sarik	yellow
Teshik	holed
Chirish	rot
Soglom	healthy

5 Results

5.1 Few-Shot Classification Performance

Table 4 and Fig. 6 summarize accuracy over 10 episodes for all six method-backbone combinations. In the 1-shot setting, fine-tuning with ResNet50 achieves the highest accuracy at $62.95\% \pm 7.28$, while fine-tuning with ViT-B/16 collapses to $24.62\% \pm 3.21$, essentially at the 25% random-chance baseline for four-way classification. Metric-learning approaches with ResNet50 reach 60.73% (Prototypical Networks) and 60.18% (k-NN (SimpleShot)) with substantially tighter confidence intervals (CI below 4.5%), indicating more stable performance across episodes. With ViT-B/16, metric-learning scores 57.55% and 58.28%, respectively, well above fine-tuning on the same backbone.

In the 5-shot setting, performance improves across all methods and confidence intervals narrow. Fine-tuning with ResNet50 remains the top performer at $81.32\% \pm 2.80$, a gain of 18.37 percentage points over its 1-shot result. Fine-tuning with ViT-B/16 improves by 34.75 points but remains 21.95 points behind ResNet50. Notably, Prototypical Networks with ViT-B/16 (75.13%) slightly surpasses Prototypical Networks with ResNet50 (73.11%) in 5-shot, reversing the ResNet50 advantage seen elsewhere; this suggests that ViT features, once enough support samples are available to form accurate prototypes, become competitive

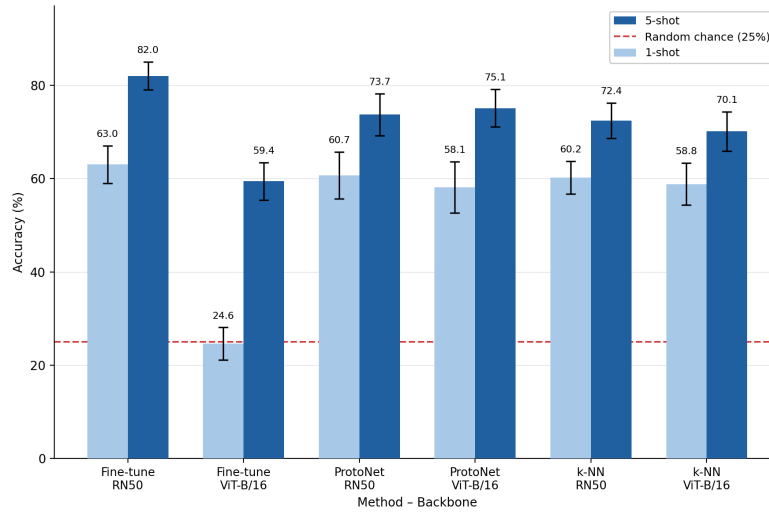


Fig. 6. 1-shot and 5-shot accuracy for all six method-backbone combinations, with 95% confidence interval error bars. The dashed red line marks the 25% random-chance baseline. Fine-tune RN50 achieves the highest 5-shot accuracy (81.32%); Fine-tune ViT-B/16 collapses to near-random at 1-shot (24.62%).

for distance-based classification even though they transfer poorly under gradient-based 1-shot fine-tuning. The k-NN gains from 1-shot to 5-shot are the most conservative (12.19 and 11.85 percentage points for ResNet50 and ViT-B/16 respectively), since frozen-feature methods refine only class prototypes or neighbor votes rather than updating representations. The highest 1-shot macro-F1 is obtained by k-NN with ResNet50 (59.30%), marginally ahead of Prototypical Networks (59.25%) and fine-tuning (58.92%); these three values overlap heavily in their confidence intervals and should be treated as statistically indistinguishable. Class-level behavior is further illustrated by the confusion matrices in Fig. 7: in the best-performing configuration (fine-tuning, ResNet50, 5-shot), correct classification rates exceed 90% for Chirish and 99% for Sarik, while Teshik reaches only 58% due to confusion with Soglom (30%), and Soglom achieves 78%; this Teshik-Soglom confusion is agronomically plausible since mild shot-hole perforations leave much of the leaf blade green and intact, making affected leaves visually similar to healthy ones. In the worst-performing configuration (fine-tuning, ViT-B/16, 1-shot), no class achieves consistently high prediction accuracy; Sarik is correctly classified only 35% of the time, and predictions are highly diffuse across all classes, reflecting the failure of linear probing on ViT-B/16 from a single support image.

Table 4. Few-shot classification performance across methods and backbones. Results are mean $\pm$ 95% CI over 10 episodes. Best value per column in **bold**; underline marks the second best.

Method	Backbone	1-Shot		5-Shot	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)
Fine-tuning	ResNet50	62.95 $\pm$ 7.28	58.92 $\pm$ 7.96	81.32 $\pm$ 2.80	80.76 $\pm$ 3.06
	ViT-B/16	24.62 $\pm$ 3.21	17.84 $\pm$ 2.83	59.37 $\pm$ 1.54	58.52 $\pm$ 1.72
Prototypical Networks	ResNet50	<u>60.73 $\pm$ 3.61</u>	<u>59.25 $\pm$ 4.19</u>	73.11 $\pm$ 1.92	72.59 $\pm$ 1.90
	ViT-B/16	57.55 $\pm$ 8.25	54.34 $\pm$ 7.94	<u>75.13 $\pm$ 4.33</u>	<u>74.62 $\pm$ 4.20</u>
k-NN (SimpleShot)	ResNet50	60.18 $\pm$ 4.35	59.30 $\pm$ 4.09	<u>72.37 $\pm$ 3.43</u>	71.85 $\pm$ 3.61
	ViT-B/16	58.28 $\pm$ 7.31	55.41 $\pm$ 6.98	70.13 $\pm$ 5.12	69.27 $\pm$ 4.86
<i>Random chance</i>	–	<i>25.00</i>	<i>25.00</i>	<i>25.00</i>	<i>25.00</i>

5.2 Zero-Shot Classification Results

Table 5 summarizes CLIP performance across prompt strategies. English prompts achieve 40.50% accuracy and 37.44% macro-F1, the best zero-shot result, but still substantially below all few-shot methods. Uzbek prompts drop accuracy to 28.75%, only 3.75 points above random chance, with macro-F1 collapsing to 18.50%, confirming that CLIP extracts almost no reliable signal from the Uzbek labels. The multilingual ensemble sits between the two at 32.00% accuracy; rather than reinforcing the English signal, averaging a well-aligned English prototype with a near-random Uzbek one dilutes the former and pulls the combined prototype away from the true visual cluster.

Table 5. Zero-shot classification results of CLIP (ViT-B/32) using different language prompt strategies on the cherry leaf disease dataset.

Language Strategy	Acc (%)	F1 (%)
English Prompts	40.50	37.44
Uzbek Prompts	28.75	18.50
Multilingual Ensemble	32.00	21.18
<i>Random chance</i>	<i>25.00</i>	<i>25.00</i>

6 Discussion

Fine-tuning with ResNet50 achieved the highest accuracy (81.32% in 5-shot, 62.95% in 1-shot), outperforming all other configurations. Across most configurations, ResNet50 outperformed ViT-B/16 by up to 38 percentage points, with the notable exception of Prototypical Networks in 5-shot, where ViT-B/16 marginally surpassed ResNet50 (75.13% vs. 73.11%), showing that convolutional

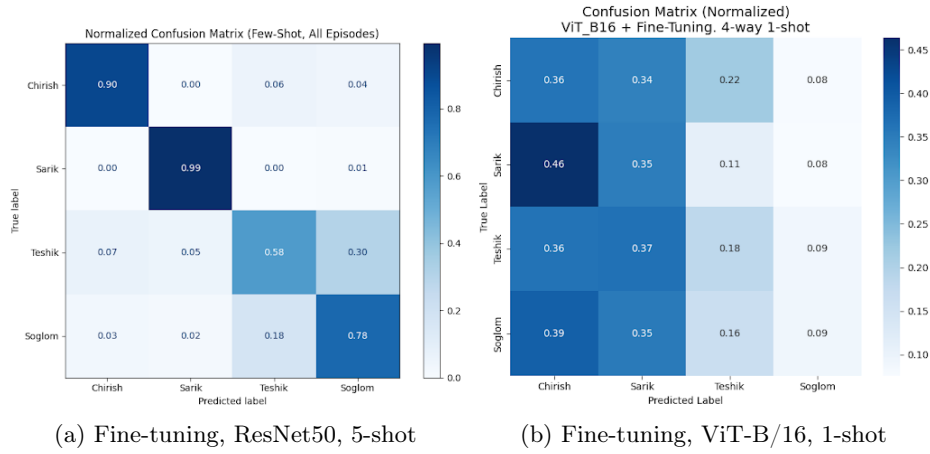


Fig. 7. Mean row-normalized confusion matrices for the best-performing (a) and worst-performing (b) configurations. Teshik-Soglom confusion in (a) is agronomically expected; the diffuse pattern in (b) indicates near-random prediction.

architectures generally maintain advantages in domain-specific few-shot learning when pre-training and target domains differ strongly. The dramatic 1-shot failure of ViT-B/16 fine-tuning (24.62%) stems from pre-training domain mismatch (ImageNet lacks sufficient agricultural textures and disease patterns), the patch-based processing of 16×16 patches potentially overlooking fine-grained disease symptoms that ResNet50’s hierarchical convolutions naturally capture, and extreme overparameterization: 86.6M parameters with only 4 support images yields a 21.6M parameter-to-sample ratio, compared to ResNet50’s more favorable 6.4M ratio. Metric-learning methods nonetheless reached 57-58% with ViT-B/16 in 1-shot, confirming that inference-only methods are more robust under very limited adaptation data. We emphasize that the ResNet50 advantage under fine-tuning reflects both architectural differences and the differing adaptation strategies (partial fine-tuning of Layer3, Layer4, and the FC head for ResNet50 versus linear-probing only for ViT-B/16, adopted due to compute constraints on a single T4 GPU); this comparison should therefore not be read as isolating architecture alone, and a matched-budget study applying the same adaptation strategy to both backbones is needed to disentangle these factors.

CLIP’s English-prompt result (40.50%) falls below even 1-shot fine-tuning (62.95%), and Uzbek prompts expose pronounced language bias (28.75%). To our knowledge, no prior study has directly measured cross-lingual bias for plant-disease classification; our result is a concrete, domain-specific instance of the English-centric bias documented in the broader vision-language literature [19, 11]. Viewed against prior work, our best result of 81.32% under 5-shot conditions on noisy field imagery with non-English labels compares favorably to the reported drop from 92.67% on laboratory images to 54.41% on field-collected images documented by Fenu and Mallocci [25], and to the 70.71% achieved on

the real-world PlantDoc dataset [26]; the gap between near-perfect curated-benchmark numbers and our realistic figures is itself a central finding.

The results reveal clear practical trade-offs: fine-tuning delivers the highest accuracy but with high variance ($\pm 7.28\%$ in 1-shot), making it sensitive to support set composition. Metric-learning methods (Prototypical Networks and k-NN (SimpleShot)) offer excellent stability (CI below 4% in 1-shot) with competitive accuracy, and the near-identical performance between them (less than 1% difference, and exactly equivalent in the 1-shot case) suggests that simple mean-pooling is sufficient for this dataset’s visually distinctive disease phenotypes. For practical deployment, fine-tuning ResNet50 is optimal when highest accuracy is required; Prototypical Networks with ResNet50 (73.11% in 5-shot) offer an excellent accuracy-stability trade-off for field deployment; ViT-B/16 fine-tuning should be avoided at 1-shot; and CLIP can support rapid prototyping but cannot replace supervised few-shot approaches where accuracy is critical.

7 Conclusion

We evaluated few-shot and zero-shot learning for cherry leaf disease classification under extreme data scarcity, using a 400-image Kyrgyzstan dataset with natural field variation and Uzbek class labels. Fine-tuning ResNet50 achieved the highest accuracy (81.32% in 5-shot), though at the cost of high variance ($\pm 7.28\%$ in 1-shot). Metric-learning methods (Prototypical Networks and k-NN (SimpleShot)) performed competitively (70–75% in 5-shot) with excellent stability (CI below 4%). ViT-B/16 underperformed under fine-tuning at 1-shot but remained competitive in 5-shot; however, since ResNet50 used partial fine-tuning while ViT-B/16 used linear probing only, this comparison should be treated as indicative rather than strictly controlled. CLIP was limited by both domain shift and language bias, reaching only 40.50% with English prompts and 28.75% with Uzbek prompts.

Limitations include the following. Computational constraints prevented a full hyperparameter search [22], and evaluation over 10 episodes leaves residual support-selection variance (1–8% CI) widest in the 1-shot fine-tuning case [23]. Labels reflect grower terminology rather than agronomist-validated diagnoses, and all data come from a single crop, region, and four categories, limiting generalizability. No augmentation was applied, likely suppressing low-shot accuracy. Because a standard random split was used without grouping by tree or farm ID, distinct leaves from the same tree or farm may appear across both the support and query sets within an episode, which may inflate reported accuracy relative to a strict tree- or farm-disjoint partitioning. Future work should validate across crops and geographies, evaluate with a farm-held-out split for unseen-farm generalization, pursue domain-adaptive pre-training of vision-language models, and conduct field validation of accuracy, latency, and treatment impact.

Acknowledgment

We thank the cherry growers of Uch-Korgon (Kadamzhay District, Batken Region, Kyrgyzstan) for granting access to their orchards and for sharing the local disease terminology used to label the dataset.

Ethics and Data Consent

All images were collected with the explicit permission of the participating growers in Uch-Korgon, who granted access to their orchards for photography. The dataset contains only images of cherry leaves; no human subjects, faces, or personally identifiable information were captured or included. Growers were informed that the images and associated Uzbek disease labels would be used for research purposes and released publicly.

Code and Data Availability

The full dataset of 400 cherry leaf images is publicly available on Mendeley Data at <https://data.mendeley.com/datasets/wp3b6pz9gc/3>. The code used for training and evaluation is not currently in a public repository. Reasonable requests for the code may be made to the corresponding author.

References

1. Bayoudh, K.: A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges. *Information Fusion* **105**, 102217 (2024)
2. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115**(3), 211–252 (2015)
3. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)* **53**(3), 1–34 (2020)
4. Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys* **55**(13s), 1–40 (2023)
5. Hughes, D., Salathé, M., et al.: An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060* (2015)
6. Geetharamani, G., Pandian, A.: Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Computers & Electrical Engineering* **76**, 323–338 (2019)
7. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems* **30** (2017)
8. Wang, Y., Chao, W.-L., Weinberger, K.Q., Van Der Maaten, L.: Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623* (2019)

9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
12. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
13. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1199–1208 (2018)
14. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135. PMLR (2017)
15. Yao, J., Tran, S.N., Garg, S., Sawyer, S.: Deep learning for plant identification and disease classification from leaf images: multi-prediction approaches. *ACM Computing Surveys* **56**(6), 1–37 (2024)
16. Ahmad, A., Saraswat, D., El Gamal, A.: A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. *Smart Agricultural Technology* **3**, 100083 (2023)
17. Bouguettaya, A., Zarzour, H., Kechida, A., Taberkit, A.M.: A survey on deep learning-based identification of plant and crop diseases from UAV-based aerial images. *Cluster Computing* **26**(2), 1297–1317 (2023)
18. Skalski, P.: Make Sense: Free to use online tool for labelling photos. <https://www.makesense.ai> (2019), last accessed 2025/11/20
19. Carlsson, F., Eisen, P., Rekathati, F., Sahlgren, M.: Cross-lingual and multilingual CLIP. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6848–6854 (2022)
20. Chen, Z., Liu, G., Zhang, B.-W., Ye, F., Yang, Q., Wu, L.: AltCLIP: Altering the language encoder in CLIP for extended language capabilities. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 8666–8682 (2023)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
22. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 (2016)
23. Landolsi, E., Kahl, F.: Tiny models from tiny data: Textual and null-text inversion for few-shot distillation. arXiv preprint arXiv:2406.03146 (2024)
24. Vishneratin, A.: NLLB-CLIP—train performant multilingual image retrieval model on a budget. arXiv preprint arXiv:2309.01859 (2023)
25. Fenu, G., Mallocci, F.M.: Evaluating impacts between laboratory and field-collected datasets for plant disease classification. *Agronomy* **12**(10), 2359 (2022)
26. Sethi, A., V, S.G., J, J.R.: Plant disease identification on real-world data using deep learning: A comparative study. In: 2021 4th International Conference on Computing and Communications Technologies (ICCCT), pp. 223–229. IEEE (2021)