scientific reports

Check for updates

OPEN

The impact of fine-tuning paradigms on unknown plant diseases recognition

Jiuqing Dong[1,2], Alvaro Fuentes[1,2], Heng Zhou[1], Yongchae Jeong[1], Sook Yoon[3]✉ & Dong Sun Park[1,2]✉

Plant diseases pose significant threats to agriculture, impacting both food safety and public health. Traditional plant disease detection systems are typically limited to recognizing disease categories included in the training dataset, rendering them ineffective against new disease types. Although out-of-distribution (OOD) detection methods have been proposed to address this issue, the impact of fine-tuning paradigms on these methods has been overlooked. This paper focuses on studying the impact of fine-tuning paradigms on the performance of detecting unknown plant diseases. Currently, fine-tuning on visual tasks is mainly divided into visual-based models and visual-language-based models. We first discuss the limitations of large-scale visual language models in this task: textual prompts are difficult to design. To avoid the side effects of textual prompts, we futher explore the effectiveness of purely visual pre-trained models for OOD detection in plant disease tasks. Specifically, we employed five publicly accessible datasets to establish benchmarks for open-set recognition, OOD detection, and few-shot learning in plant disease recognition. Additionally, we comprehensively compared various OOD detection methods, fine-tuning paradigms, and factors affecting OOD detection performance, such as sample quantity. The results show that visual prompt tuning outperforms fully fine-tuning and linear probe tuning in out-of-distribution detection performance, especially in the few-shot scenarios. Notably, the max-logit-based on visual prompt tuning achieves an AUROC score of 94.8% in the 8-shot setting, which is nearly comparable to the method of fully fine-tuning on the full dataset (95.2%), which implies that an appropriate fine-tuning paradigm can directly improve OOD detection performance. Finally, we visualized the prediction distributions of different OOD detection methods and discussed the selection of thresholds. Overall, this work lays the foundation for unknown plant disease recognition, providing strong support for the security and reliability of plant disease recognition systems. We will release our code at https://github.com/JiuqingDong/PDOOD to further advance this field.

Keywords Few-shot learning, Open-set recognition, Out-of-distribution detection, Plant disease recognition, Visual prompt

Plant disease recognition is critical for farmers and agricultural researchers, as diseases can rapidly spread across crops, causing substantial yield and economic losses. Annually, plant diseases result in an estimated global economic cost of USD 220 billion, primarily due to bacteria, fungi, nematodes, and viruses[1,2]. To mitigate these impacts, the application of deep learning in general computer vision tasks has been extended to plant disease recognition, demonstrating significant potential[3,4].

Deep learning methods learn the feature representations by utilizing multiple processing layers such as perceptrons, convolutional layers, or transformers. This end-to-end approach is particularly advantageous for its effectiveness in capturing complex patterns and features directly from the raw data. Among these methods, Convolutional Neural Networks (CNNs) have been particularly transformative, eliminating the need for manual feature extraction from images. We have witnessed impressive achievements in the application of deep learning for plant disease detection, where the classification accuracy often exceeds 90%[5–8]. However, most existing studies focus on fixed disease categories of specific species with all available annotations during the training phase. In

[1]Department of Electronic Engineering, Jeonbuk National University, Jeonju 54896, South Korea. [2]Core Research Institute of Intelligent Robots, Jeonbuk National University, Jeonju 54896, South Korea. [3]Department of Computer Engineering, Mokpo National University, Muan-gun 58554, South Korea. ✉email: syoon@mokpo.ac.kr; dspark@jbnu.ac.kr

this case, the model is more likely to classify a suspicious sample as one of the categories it has already learned rather than indicating the presence of an abnormal disease type, which adds potential risks to the system[9].

In an open-world scenario, the assumption that test set categories mirror training set categories is often unrealistic in practical applications, particularly in complex plant diseases. For instance, unforeseen diseases and pests can emerge during plant growth cycle. We demonstrate examples of the aforementioned potential risks, as shown in Fig. 1: The model trained on healthy leaves from 12 different species classifies disease samples as one of the known categories with a high confidence score. We argue that a reliable model should assign lower confidence scores to these samples, indicating that they belong to unknown categories not in the training dataset[10]. This attribute is crucial for the safety and reliability in plant disease recognition systems. This challenge is referred to as out-of-distribution (OOD) detection or open-set recognition (OSR)[11,12]. Although there may be terminological differences, OOD detection and OSR (sometimes referred to as "novelty detection") essentially pursue the same goal: detecting and excluding unknown samples. The differences in experimental settings between OOD detection and OSR are not the focus of this paper. Therefore, we uniformly use "OOD detection" to avoid terminological confusion.

Recent OOD detection methods focused on large vision-language models[13–15]. For example, Ming et al.[13] proposed a method to identify unknown samples, called Maximum Concept Matching (MCM). MCM utilizes aligned visual and semantic information, using conceptual features as classification weights for zero-shot predictions during the inference process. The performance of MCM even surpasses that of methods based on fine-tuning[11,16–18], depending on the generalization capabilities of large-scale vision-language models. However, we argue that employing large-scale vision-language models like CLIP directly for OOD detection in plant disease recognition proves impractical. As shown in Fig. 2, we attempt to compare the common datasets settings from[13] and plant disease datasets in the OOD detection task by using the CLIP[19] model. It can be observed that CLIP can better separate ID and OOD data in common datasets rather than plant disease datasets. We further compared the performance of large-scale vision-language models in OOD detection for plant disease under various language prompt settings. As shown in Table 1, different language prompts have varied effects on the performance of the vision-language models, but none perform well. The failure of this approach stems from two key factors: **(i).** The pre-training data of CLIP and the training data for the downstream task of plant disease detection have a significant domain gap. Plant disease recognition involves a more fine-grained classification task as leaf samples can be highly similar, making it difficult for the vision language model to distinguish between different diseases without training; **(ii).** Language prompts are crucial for vision language models, but it is challenging to design efficient prompts for plant disease. Despite these limitations, the study by Ming et al.[13] has opened new avenues

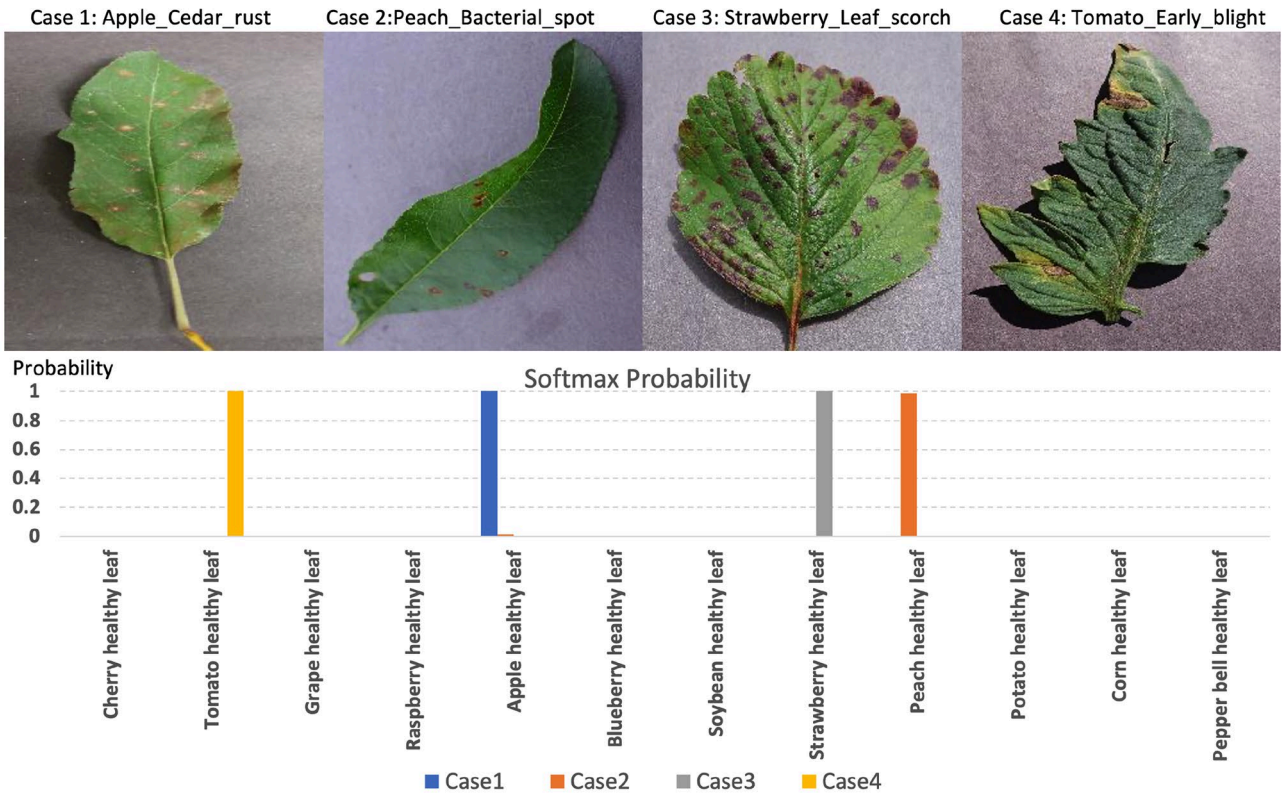


Figure 1. The risks faced by the existing model. We provide four case with the softmax probability distribution. The model classifies four diseased leaves as one of the training categories with a high confidence score. Note that the model was pre-trained on ImageNet-21k and fine-tuned on healthy leaves from 12 different species. The horizontal axis shows the names of the training categories.

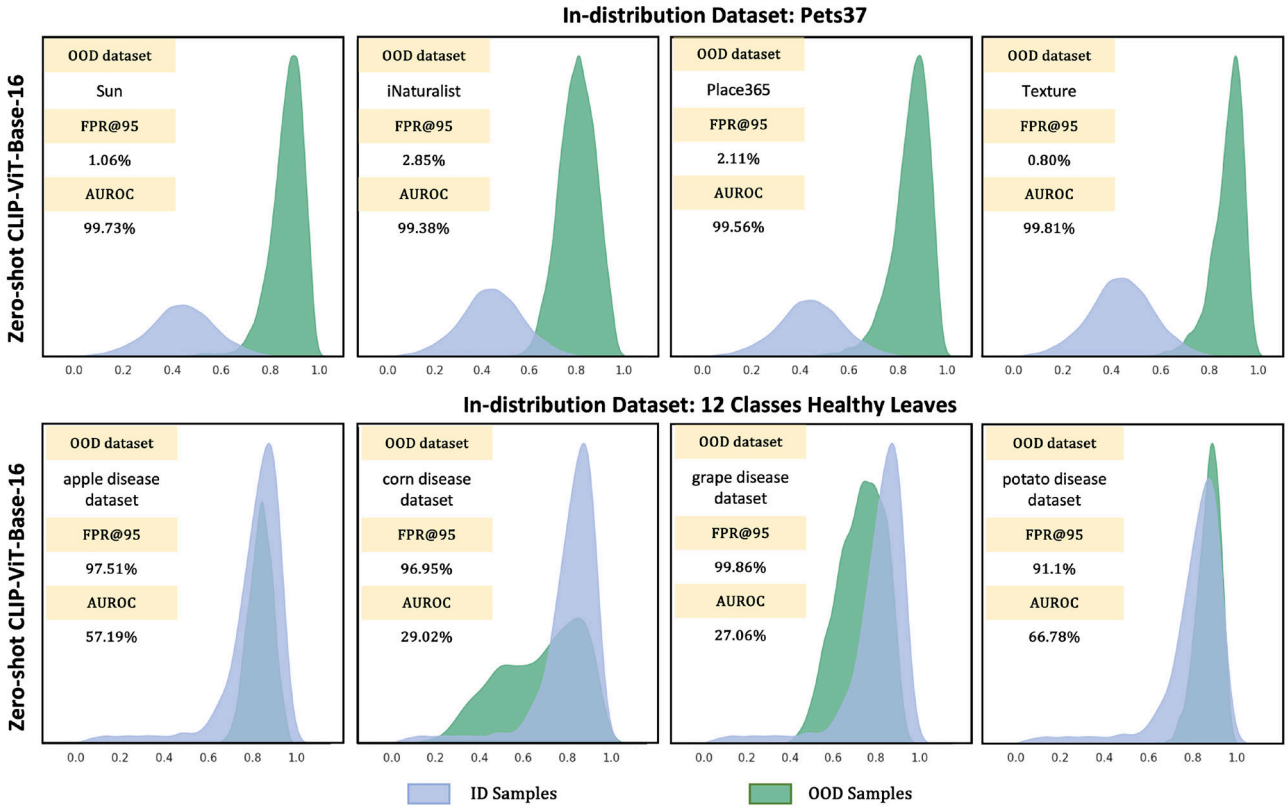


Figure 2. Kernel density estimation plot for zero-shot out-of-distribution detection using a visual-language model CLIP. CLIP can better separate ID and OOD data in common datasets (Top) rather than plant disease datasets (Bottom). Note that the language prompt type is:"a photo of a Class names". The model used for training is CLIP-ViT-base-16. Uncertainty scores are calculated using maximum concept matching (MCM)[13]. OOD datasets and evaluation metrics are added to the top left corner of each subplot.

Pre-trained model	Language prompt	FPR@95↓	AUROC↑	AUPR↑
CLIP-ViT-base-16	{Class names}	95.52	49.23	49.66
	This is a photo of {Class names}	89.61	52.61	50.83
	a photo of a {Class names}	91.41	52.48	52.12
CLIP-ViT-large-14	{Class names}	95.56	49.02	55.92
	This is a photo of {Class names}	95.63	45.62	52.19
	a photo of a {Class names}	97.51	53.03	59.54

Table 1. Results of Zero-Shot Out-of-Distribution Detection by using different language prompts. The class names for in-distribution data are shown in Fig. 1. We present the average test results on six out-of-distribution datasets. ↓ indicates that lower values are preferable, and conversely for ↑.

in OOD detection by leveraging the rich visual-semantic information of large-scale models. This approach has inspired us to explore the potential of large-scale pre-training models.

Language prompts are crucial for visual language models, but disease categories are not as common and straightforward as 'cat' or 'dog'. Although some text-prompt-based fine-tuning methods[20,21] can automatically generate language prompts for the text branch, we still observe some limitations, such as ID accuracy may be lower than that of unimodal models. We will elaborate on these details further in the discussion. Due to the difficulty in designing language prompts specifically for plant diseases, we rethink a question: How can we leverage unimodal visual models to achieve the out-of-distribution detection for plant diseases? To our knowledge, we are the first to comprehensively discuss the impact of fine-tuning paradigms on the performance of Open Set Recognition (OSR), Out-of-Distribution (OOD) detection, and few-shot learning tasks in plant disease recognition. Our main contributions are as follows:

- We experimentally demonstrate that zero-shot visual language models perform poorly in plant diseases' fine-grained OOD detection tasks. To avoid the cumbersome task of designing language prompts, we first

investigate the impact of different fine-tuning paradigms on the pure vision pre-trained model in plant disease OOD detection.
- We are the first to establish a comprehensive benchmark of unknown plant disease recognition. Our benchmark covers results of fully fine-tuning (FFT), linear probe tuning (LPT), and visual prompt tuning (VPT) across five public datasets under various experimental conditions. To our knowledge, there is currently no related work studying the impact of fine-tuning paradigms on out-of-distribution detection for plant diseases.
- We employ various OOD detection methods to evaluate the effectiveness of different fine-tuning paradigms. The research results demonstrate the promising prospects of visual prompts in open-set detection, OOD detection, and few-shot learning tasks for plant disease classification, providing robust support for the safety and reliability of plant disease recognition systems.

The organization of the paper is as follows: Section "Related work" states the related work including post-hoc OOD detection methods and fine-tuning paradigms. Secion "Materials and methods" introduces the problem statement and plant disease datasets in our benchmark. Additionally, we proposed our framework, a brief overview of the current popular fine-tuning paradigms, post-hoc OOD detection methodologies, and evaluation metrics. Section "Experiments and results" presents the experimental results under open-set, OOD, and few-shot learning settings. Finally, we discuss the advances and limitations of this work and potential future directions.

Related work

Post-hoc-based OOD detection

In the latest review on OOD detection, Yang et al.[22] highlight the advantages of post-hoc OOD detection methods, noting their ease of integration without altering training procedures and objectives. This characteristic is crucial in real-world applications where indirect retraining costs can be significant. An early approach, the maximum softmax probability (MSP) method[11], assumes that ID samples typically have higher maximum softmax probabilities than misclassified or OOD samples. This method also involves estimating uncertainty scores through information entropy derived from softmax probability distributions. Another seminal work[23] enhances the distinction between in-distribution (ID) and OOD samples through temperature scaling and input perturbation. Hendrycks et al.[24] argue that reliance on softmax confidence scores may lead to overconfidence in the posterior distribution of OOD data, and propose using maximum logits (ML) for more effective OOD detection. In contrast, Liu et al.[25] demonstrate that the energy score, which aligns with the input's probability density, is less prone to overconfidence. They advocate that energy can function both as a scoring mechanism for pre-trained neural classifiers and as a trainable loss function to specifically tailor the energy surface for OOD detection. Lin et al.[26] theoretically support this approach, suggesting that lower energy scores indicate ID samples and higher scores suggest OOD samples, equating energy scores with uncertainty measures.

A recent study by Ming et al.[13] introduces a novel zero-shot OOD detection approach using the large-scale vision-language model CLIP[19]. Test images and ID class labels are embedded into respective visual and text encoders, generating visual and textual features. Cosine similarity between these features is then used as logits, and maximum softmax probability is employed for OOD detection, a method termed "Maximum Concept Matching (MCM)." Miyai et al.[15] proposed using local regularization techniques and fine-tuning the CLIP model to enhance the out-of-distribution detection performance of MCM. The effectiveness of this technique is contingent upon the generalization capabilities of large-scale vision-language models.

Fine-tuning paradigm

Hendrycks et al.[27] contend that pre-training significantly boosts a model's adversarial robustness, outperforming state-of-the-art methods in robustness and uncertainty tasks. Recent large-scale pre-training models like CLIP[19] and SAM[28] exhibit impressive stability in zero-shot tasks. Consequently, we argue that transfer learning is a promising strategy that can maximize the utility of pre-trained models under limited training data. To effectively harness the robustness of large-scale pre-trained models, we investigated three different transfer learning strategies: fully fine-tuning (FFT), linear probe tuning (LPT)[29], and visual prompt tuning (VPT)[30]. We briefly summarize these three fine-tuning paradigms as follows: FFT involves updating all parameters for extensive adaptation, though it requires more resources. Linear probe tuning only updates classification head parameters, keeping the backbone frozen. LPT is particularly suited for few-shot learning as it helps prevent overfitting on small training sets; Visual prompt tuning introduces a small number of trainable tokens into the input space, keeping the backbone intact. Following Jia et al.[30], we incorporate ten learnable prompt tokens in each transformer layer.

Materials and methods

This section outlines the datasets, including their splits and experimental settings. We also discuss the vision transformer[31], a popular feature extractor in recent research. However, we will not delve into its specific parameters, layers, or self-attention mechanism. Instead, our focus will be on the advantages of the visual sensor and the rationale for its selection. Lastly, we will describe various fine-tuning methods, OOD detection methods, and evaluation metrics used in our study.

Problem statement

In this section, we define the anomaly detection problem. The training set is $D^{train} = \{x_i, y_i\}_i^N, i \in N$, where x, y and N denote the sample, label, and the number of images. We define the set of known classes as $K = \{k_1, k_2, k_3, ...k_t\}$, so $y_i \in K$. In particular, for few-shot settings we have training set $D^{train} = \{x_i, y_i\}_{i=1}^{M*k_t}, M \in \{2, 4, 8, 16\}$, where M and k_t denote the number of training samples of each known class and the number of classes. We assume that there

ID	Cotton Class name (Images)	Mango Class name (Images)	Strawberry Class name (Images)	Tomato Class name (Images)
1	Healthy (800)	Healthy (500)	Healthy (456)	Healthy (1591)
2	Powdery mildew (800)	Sooty Mould (500)	Powdery mildew leaf (533)	Early blight (1000)
3	Target spot (800)	Anthracnose (500)	Anthracnose fruit rot (97)	Leaf Mold (952)
4	Aphids (800)	Powdery Mildew (500)	Leaf spot (615)	Spider mites (1676)
5	Bacterial blight (800)	Bacterial Canker (500)	Powdery mildew fruit (135)	Septoria leaf spot (1771)
6	Army worm (800)	Die Back (500)	Blossom blight (208)	Mosaic virus (373)
7	–	Cutting Weevil (500)	Angular leafspot (435)	Bacterial spot (2127)
8	–	Gall Midge (500)	Gray mold (477)	Late blight (1909)
9	–	–	–	Yellow Leaf Curl Virus (5357)
10	–	–	–	Target Spot (1404)

Table 2. Classes information of the dataset. We have assigned an ID to each category to facilitate the representation of known and unknown classes.

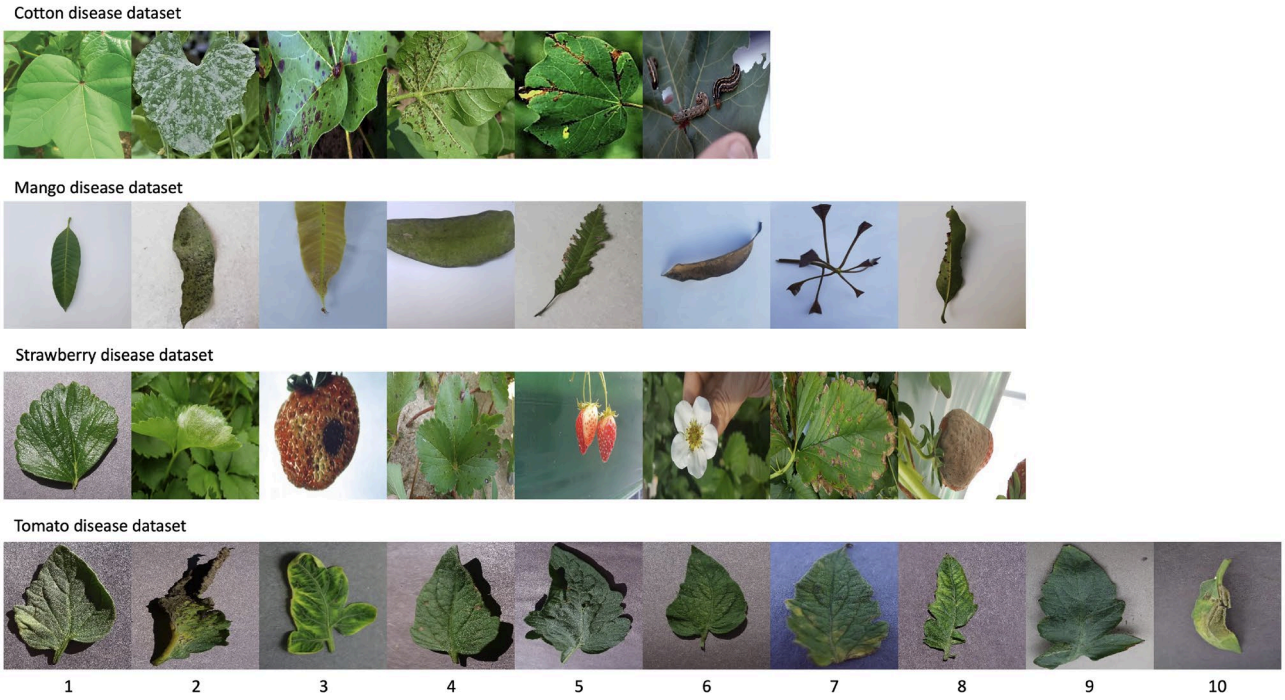


Figure 3. Dataset examples used for open-set recognition. Class numbers are provided at the bottom. Please refer to Table 2 for class names.

exists a set of unknown classes $U = \{k_{t+1}, ...\}$, which the model does not witness during training but may encounter during inference, and $K \cap U = \emptyset$. That means unknown samples should not have labels that overlap with the training data. We can define anomaly detection as a binary classification problem which is formalized as Eq. 1:

$$Decision_{\gamma}(x_i) = \begin{cases} Unknown\ Class & S(x_i) > \gamma \\ Known\ Class & S(x_i) \leq \gamma \end{cases}, \tag{1}$$

where a higher score $S(x_i)$ for a sample x_i indicates higher uncertainty. A sample with a score greater than the threshold γ will be classified as unknown classes, and vice versa.

Datasets

For our experiments on open-set recognition, we employ the Cotton[32], Mango[33], Strawberry[34], and Tomato[35] disease datasets. We illustrate samples from these datasets in Fig. 3 and provide the detailed training and testing splits in Table 2. Additionally, the Plant Village dataset[35] is utilized for our OOD detection and few-shot OOD detection experiments. In addition to the ID categories shown in Fig. 1, we present the OOD categories in Table 5 of the experimental section. Due to the extensive range of categories within the Plant Village dataset, they are not all displayed here. We confirm that all aspects of our study, including both experimental research and field studies on plants, have been conducted in strict accordance with the relevant guidelines and legislation. This

compliance covers institutional protocols as well as national and international regulations about plant research. For further details, please refer to our repository.

Cotton disease dataset[32] comprises five plant diseases, including Aphids, Army Worm, Bacterial Blight, Powdery Mildew, and Target Spot. Its primary focus is on leaf diseases, with no images of diseases affecting stems, buds, flowers, or bolls. The dataset features a balanced class distribution, with around 800 images per category, and was collected in real-world conditions, Besides, it also provides 800 images for healthy leaves.

Mango leaves disease dataset[33] compiled by Ahmed et al., is a comprehensive collection of 4000 mango leaf images, each with a resolution of 240x320 pixels. The dataset encompasses seven specific mango leaf diseases and healthy leaves. Each disease category contains roughly 500 images, ensuring a balanced distribution across the eight classes. The images, mainly captured using mobile phone cameras, originate from four mango orchards in Bangladesh.

Strawberry disease dataset[34] released by Afzaal et al., this dataset includes 2500 images that capture various strawberry diseases. The data were collected using camera-equipped mobile phones in both real-field and greenhouse settings, mainly across multiple greenhouses in South Korea. This dataset, encompassing the early, middle, and late stages of the diseases, was designed to enhance disease detection and segmentation. To ensure consistency with other datasets, we have supplemented it with images of healthy strawberry leaves from the Plant Village dataset[35].

Tomato disease dataset[35] focuses on tomato diseases. It includes ten categories of tomato leaves, encompassing nine disease types and one healthy leaf category. This dataset is notable for its imbalanced sample distribution, ranging from 300 to 5000 samples per category, which adds complexity to the analysis. For comprehensive evaluation, we have utilized color, grayscale, and segmented images.

Plant village dataset[35] is a vast collection of 54,309 images, covering 14 crop species and a wide range of diseases, including fungal, bacterial, oomycete, viral, and mite-induced diseases. It also features healthy leaves for twelve crop species. For our study, we used images of 12 types of healthy leaves as an in-distribution (ID) dataset and constructed six out-of-distribution (OOD) datasets based on species categorization: apple (3 types), corn (3), grape (3), potato (2), tomato (9), and others (6). We also evaluated OOD detection performance under a few-shot learning setting using this partitioning approach.

Overview of framework

We present an overview of the framework for this study in Fig. 4. The post-hoc out-of-distribution detection method consists of two steps. The first step involves training or fine-tuning the model on a training set. In the second step, post-hoc OOD detection methods are deployed to obtain uncertainty scores, such as those based on energy and maximum softmax probability.

In our study, we employed the ViT-base model as a feature extractor to assess the effectiveness of various fine-tuning paradigms in open-set recognition (OSR), out-of-distribution (OOD) detection, and few-shot OOD detection. One of the significant advantages of ViT is its ability to achieve remarkable performance on large-scale image datasets with minimal architectural modifications. For example, a Transformer model trained on textual data can be directly used for fine-tuning visual tasks. It can be easily fine-tuned for specific tasks such as object

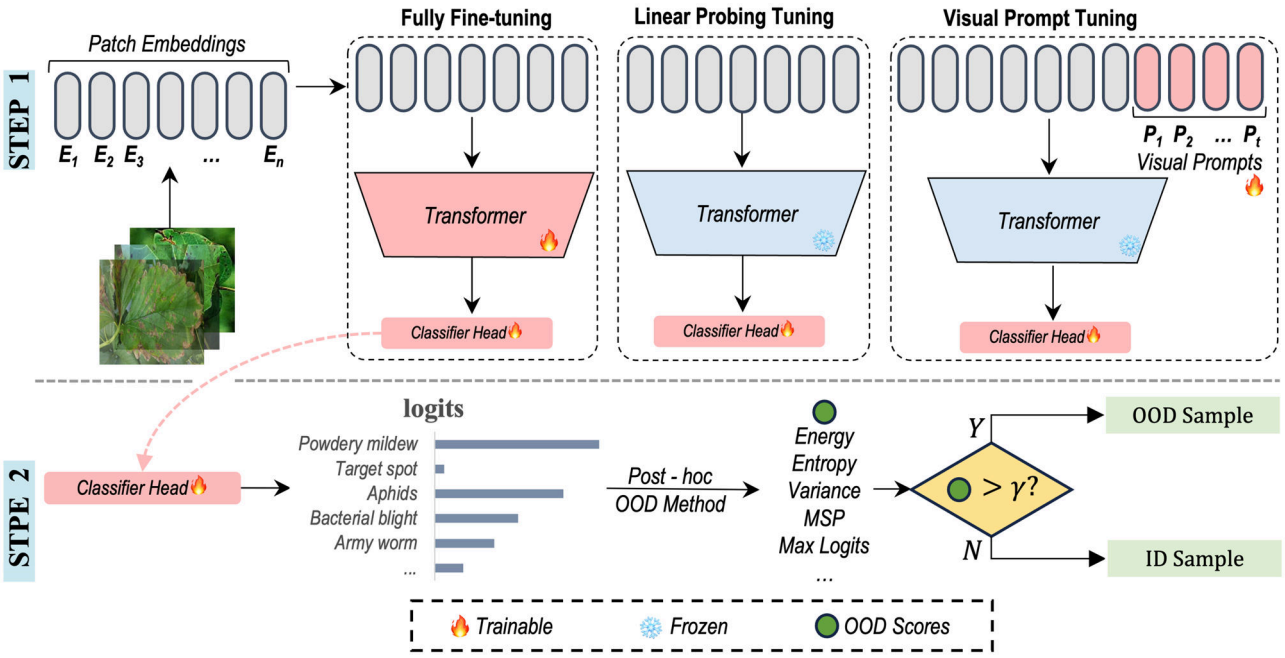


Figure 4. The architecture of three fine-tuning paradigms for OOD detection. Step 1 compares three fine-tuning paradigms for ViT model, where visual prompts fine-tune the model by adding a set of learnable tokens to the input space. Step 2 indicates the pipeline of post-hoc OOD detection methods.

detection[36], key-point detection localization[37], and image segmentation[28], further demonstrating its flexibility and ability to generalize across different vision tasks.

In Step 1, we use ViT-B/16, pre-trained on ImageNet-21k, as our pre-trained model. To address the problem of OOD detection, we explored two traditional fine-tuning paradigms (FFT and LPT), and an efficient fine-tuning paradigm (VPT). The three different fine-tuning paradigms are illustrated in Fig. 4. It can be observed that all of them involve fine-tuning the classifier head. In Step 2, the logits output by the classifier head are transformed into uncertainty scores $S(x_i)$ using different post-processing out-of-distribution (OOD) detection methods. If $S(x_i)$ exceeds a threshold γ, the sample is treated as an OOD sample, thus achieving out-of-distribution detection.

OOD detection methods

The classification head can be considered as a feature mapping that aims to map the input image's features $F \in R^d$ to the label space $L \in R^c$, where c represents the number of classes in ID dataset, and L represents the logic values for each class. Post-hoc methods for OOD detection estimate the distributions of ID and OOD data by processing these logic values as an uncertainty score, thereby separating them. The advantages of post-hoc methods lie in their ease of use and the fact that they do not require any modification of the training process and loss function. Consistent with the OOD detection methods used in[13], in this paper, we employed five commonly used post-hoc processing methods: energy[25], entropy[11], variance[38], maximum softmax probability (MSP)[11], and maximum logits (ML)[24]. The conversion formula from logical values to uncertainty scores is shown in Eq. 2 to Eq. 6.

$$Energy = -log \sum_{j=1}^{K} z_j/T \tag{2}$$

$$Entropy = Entropy\left(\frac{e^{z_i}/T}{\sum_{j=1}^{K} e^{z_j}/T}\right) \tag{3}$$

$$Variance = -Variance\left(\frac{e^{z_i}/T}{\sum_{j=1}^{K} e^{z_j}/T}\right) \tag{4}$$

$$MSP = -Max\left(\frac{e^{z_i}/T}{\sum_{j=1}^{K} e^{z_j}/T}\right) \tag{5}$$

$$Max - Logits = -Max\left(\frac{z_i/T}{\sum_{j=1}^{K} z_j/T}\right) \tag{6}$$

where z_i denotes the logits of class i, and T denotes the temperature scaling factor. In this paper, we used $T = 1$ as default.

Evaluation metrics

FPR@95[9]: FPR@95 can be interpreted as the probability that a negative (out-of-distribution) example is misclassified as positive (in-distribution) when the true positive rate (TPR) is as high as 95%. The true positive rate can be computed by TPR = TP/(TP + FN), where TP and FN denote true positives and false negatives, respectively. The false positive rate (FPR) can be computed by FPR = FP/(FP + TN), where FP and TN denote false positives and true negatives, respectively.

The area under the receiver operating characteristic curve (AUROC)[39]: By treating ID data as positive and OOD data as negative, various thresholds can be applied to generate a range of true positive rates (TPR) and false-positive rates (FPR). From these rates, we can calculate AUROC.

The area under the precision-recall curve (AUPR)[39]: Using the precision and recall values, we can compute metrics of AUPR. Please note that for AUROC and AUPR, higher values indicate better OOD detection performance, while a lower FPR@95 value indicates better OOD detection performance.

In-Distribution Accuracy (ID Acc.)[40]: OOD detection and open-set recognition also require evaluating the model's performance on ID samples. Therefore, we use Accuracy as the evaluation metric for ID samples.

Experiments and results

Implementation details

Our experiments used a pre-trained Vision Transformer (ViT-base-16) model on the ImageNet-21k dataset[41] as the feature extractor. We assessed the effectiveness of three different fine-tuning strategies-fully fine-tuning, linear probe tuning, and visual prompt fine-tuning-in the context of OSR, OOD detection, and few-shot OOD detection for plant diseases. The specific class indices for the datasets are detailed in Table 2. These indices were crucial for distinguishing between known and unknown classes in open-set recognition tasks.

We employed ten learnable visual prompts for the visual prompt fine-tuning method, which is the default setting in[30]. An advantage of our method is its modest computational resource requirement. All experiments were conducted on a single Nvidia RTX 3090 GPU. We used PyTorch version 1.10.0 as our training framework. The uncertainty scores and distribution analyses were computed using the scikit-learn and numpy libraries. We

searched for the optimal learning rate and weight decay within a specific range to explore the best ID accuracy. All experimental results are reported based on the optimal ID accuracy.

Open-set recognition settings

Open-set recognition (OSR) is one of the most related to OOD detection. OSR typically involves a single multi-class dataset, where some categories are used as known classes for training. In contrast, the remaining categories are treated as unknown samples and included in the test set. Table 3 provides detailed experimental setups along with the corresponding divisions between known and unknown classes. We first evaluate the ID accuracy in these different experiment settings. The results show that the three fine-tuning methods achieved nearly 100% ID accuracy across different settings for cotton, mango, and strawberry disease datasets, as shown in Fig. 5. Furthermore, we evaluated the performance of three fine-tuning paradigms in OSR settings using three datasets: cotton, mango, and strawberry. For example, there are five data splits for the cotton disease dataset, and we provide average results across these five experimental settings. Visual prompt tuning significantly outperformed both fully fine-tuning and linear probe tuning in OSR tasks. Our experiments indicate that no single method is consistently superior across different benchmarks, and performance rankings can vary significantly between datasets. For example, the OOD detection method based on maximum logits is more effective with visual prompt tuning, while the method based on maximum softmax probability performs best with fully fine-tuning.

In addition, we also evaluated these methods on three versions of the Tomato disease datasets, including color, grayscale, and segmented versions. We present the dataset partitioning settings in Table 4. For a fair comparison, the image numbering remains consistent across all three versions. We present only the average results for all experimental settings to streamline the presentation and enhance readability. Regardless of the fine-tuning paradigm, the model performs better on the original color dataset, which contains more information. In contrast,

	Cotton disease dataset		Mango disease dataset		Strawberry disease dataset				
Experiment no.	**Known classes**	**Unknown classes**	**Known classes**	**Unknown classes**	**Known classes**	**Unknown classes**			
Plant disease dataset splits									
1	1,2	3,4,5,6	1,2	3,4,5,6,7,8	1,2	3,4,5,6,7,8			
2	1,2,3	4,5,6	1,2,3	4,5,6,7,8	1,2,3	4,5,6,7,8			
3	1,2,3,4	5,6	1,2,3,4	5,6,7,8	1,2,3,4	5,6,7,8			
4	1,2	5,6	1,2,3,4,5	6,7,8	1,2,3,4,5	6,7,8			
5	1,2,3	5,6	1,2,3,4,5,6	7,8	1,2,3,4,5,6	7,8			
6	–	–	1,2	7,8	1,2	7,8			
7	–	–	1,2,3	7,8	1,2,3	7,8			
8	–	–	1,2,3,4	7,8	1,2,3,4	7,8			
9	–	–	1,2,3,4,5	7,8	1,2,3,4,5	7,8			
	VPT			**LPT**			**FFT**		
Method	**FPR@95↓**	**AUROC↑**	**AUPR↑**	**FPR@95↓**	**AUROC↑**	**AUPR↑**	**FPR@95↓**	**AUROC↑**	**AUPR↑**
Cotton disease dataset									
Energy	**24.10**	**93.20**	**92.44**	70.78	84.82	88.05	40.96	88.68	87.24
Entropy	**23.01**	**93.66**	**92.93**	63.52	83.23	79.87	26.20	92.57	91.52
Variance	**23.03**	**93.67**	**92.95**	72.63	70.28	65.57	28.61	88.79	82.84
MSP	**22.97**	**94.45**	**92.95**	72.63	70.10	65.44	29.84	88.22	81.86
ML	**26.10**	**93.21**	**92.45**	70.31	84.84	88.06	40.85	88.70	87.25
Mango disease dataset									
Energy	**12.60**	**97.05**	**94.89**	39.42	87.48	84.96	21.34	92.62	89.94
Entropy	**12.94**	**96.47**	**95.38**	57.43	82.82	80.68	15.33	94.90	91.72
Variance	**12.89**	**96.41**	**95.13**	69.76	69.84	66.89	15.92	93.93	88.38
MSP	**12.89**	**96.34**	**95.04**	69.82	64.99	65.56	17.29	93.66	88.01
ML	**12.57**	**96.56**	**95.06**	39.36	87.48	84.96	21.28	92.64	89.95
Strawberry disease dataset									
Energy	**18.40**	**93.15**	**91.45**	42.50	89.72	89.27	30.92	92.16	91.69
Entropy	**18.45**	94.31	92.26	52.27	85.69	86.05	19.81	**95.97**	**95.94**
Variance	**18.72**	**94.33**	**92.27**	61.55	76.34	75.89	21.14	94.19	90.73
MSP	**18.87**	**94.33**	**92.27**	61.95	72.33	73.15	23.18	93.60	89.76
ML	**18.42**	**93.15**	**91.45**	42.30	89.74	89.29	30.85	92.17	91.70

Table 3. Open-set recognition experiments settings and results on cotton, mango, and strawberry disease datasets. A set of experiments were conducted for each dataset. We provide divisions between known and unknown classes for each experimental setting. The results are based on the average across all experimental settings for each dataset. Bold indicates the best performance.

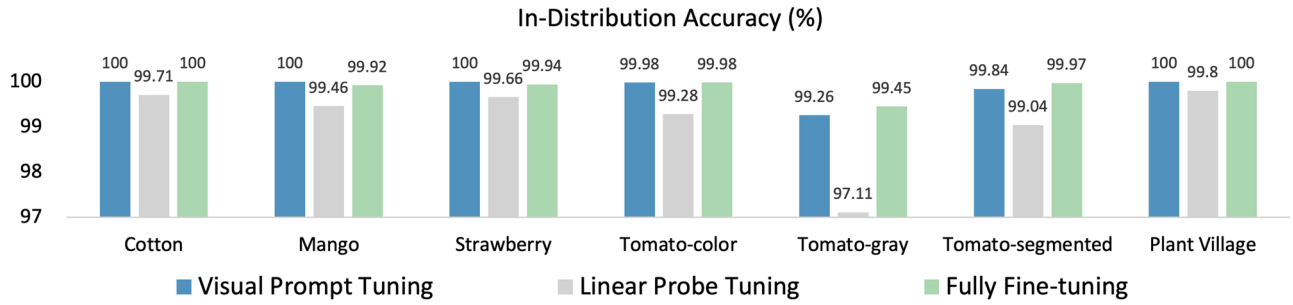


Figure 5. Results of ID accuracy on five datasets.

Experiment no.	Known classes	Unknown classes	Experiment no.	Known classes	Unknown classes				
Tomato disease dataset splits									
1	1,2	3,4,5,6,7,8,9,10	7	1,2	8,9,10				
2	1,2,3	4,5,6,7,8,9,10	8	1,2,3	8,9,10				
3	1,2,3,4	5,6,7,8,9,10	9	1,2,3,4	8,9,10				
4	1,2,3,4,5	6,7,8,9,10	10	1,2,3,4,5	8,9,10				
5	1,2,3,4,5,6	7,8,9,10	11	1,2,3,4,5,6	8,9,10				
6	1,2,3,4,5,6,7	8,9,10	-	–	–				
	VPT			**LPT**			**FFT**		
Method	**FPR@95↓**	**AUROC↑**	**AUPR↑**	**FPR@95↓**	**AUROC↑**	**AUPR↑**	**FPR@95↓**	**AUROC↑**	**AUPR↑**
Color version									
Energy	**14.42**	**95.85**	**93.76**	46.87	86.55	86.26	20.89	94.44	92.82
Entropy	**15.02**	95.64	93.61	45.71	86.02	85.29	15.84	**95.92**	**95.31**
Variance	**15.05**	**95.61**	**93.60**	48.68	83.90	81.75	16.35	94.44	89.21
MSP	**15.05**	**95.60**	**93.60**	49.08	83.61	81.58	16.74	92.99	88.33
ML	**14.43**	**95.85**	**93.77**	45.99	86.69	86.31	20.88	94.47	92.82
Gray version									
Energy	**37.45**	**87.97**	**86.40**	66.60	81.97	83.25	42.24	86.91	85.69
Entropy	**40.00**	87.61	86.38	59.89	84.39	84.40	40.90	**88.58**	**86.91**
Variance	40.83	**87.50**	**86.34**	60.37	83.20	83.34	**40.81**	84.93	81.42
MSP	40.89	**87.47**	**86.32**	61.05	83.06	83.28	**40.83**	84.69	83.01
ML	**37.44**	**87.96**	**86.40**	65.26	82.68	83.42	42.16	86.91	85.69
Segmented version									
Energy	**21.57**	**94.09**	**93.62**	48.17	87.28	86.65	24.90	93.84	92.10
Entropy	22.36	94.01	93.65	47.22	88.52	85.93	**19.81**	**95.38**	**94.55**
Variance	22.56	**93.98**	**93.63**	53.28	84.17	81.68	**20.69**	90.73	86.13
MSP	**22.57**	**93.97**	**93.63**	53.91	83.96	81.55	24.34	89.83	84.90
ML	**21.50**	**94.09**	**93.62**	47.35	88.38	86.72	24.90	93.84	92.10

Table 4. Open-set recognition experiments settings and results (%) on tomato disease dataset. 11 experiments are conducted for the tomato disease dataset. We provide divisions between known and unknown classes for each experimental setting. The results are based on the average across all experimental settings for each dataset. Bold indicates the best performance.

the model's performance significantly declined when trained on grayscale images, especially in evaluating OOD metrics. For example, compared to the color dataset, the visual prompt method's AURPC dropped from 95.85% to 87.97% in the energy-based OOD detection method. The primary reasons for this decline are two-fold: firstly, the decrease in ID accuracy affects the model's performance in uncertainty tasks; secondly, color is a crucial factor in distinguishing certain diseases, and the absence of color information in grayscale images means that some OOD disease categories may appear visually similar to ID disease categories.

We also examined the impact of background on recognition performance by evaluating the three fine-tuning paradigms on segmented versions of the dataset, where backgrounds were replaced with black or white, isolating leaf images. The results show a slight decline in model performance compared to the original color images, regardless of the fine-tuning paradigm. Even so, visual prompt tuning consistently exhibits excellent performance. Note that the results in Table 3, Table 4, and Fig. 5 are the average results of all experiments for each dataset. More detailed results are also available in our code repository for further reference.

Dataset type	Plants (Number of Classes)								
Plant village dataset splits									
ID	Cherry (1), Tomato (1), Grape (1), Raspberry (1), Apple (1), Blueberry (1), Soybean (1), Strawberry (1), Peach (1), Potato (1), Corn (1), Bell Pepper (1)								
OOD	Apple (3)								
OOD	Corn (3)								
OOD	Grape (3)								
OOD	Potato (2)								
OOD	Tomato (9)								
OOD	Cherry (1), Orange (1), Peach (1), Bell Pepper (1), Squash (1), Strawberry (1)								
	VPT			**LPT**			**FFT**		
Method	**FPR@95↓**	**AUROC↑**	**AUPR↑**	**FPR@95↓**	**AUROC↑**	**AUPR↑**	**FPR@95↓**	**AUROC↑**	**AUPR↑**
Plant village dataset									
Energy	**10.69**	**97.73**	**96.43**	14.53	96.15	94.56	13.60	95.29	86.89
Entropy	10.11	**97.64**	**96.46**	21.06	95.08	93.66	**8.57**	97.33	92.09
Variance	10.35	**97.60**	**96.43**	23.14	94.82	93.44	**8.33**	97.08	90.90
MSP	10.35	**97.60**	**98.10**	23.19	94.78	93.42	**8.25**	96.94	90.13
ML	**7.40**	**98.25**	**96.42**	15.09	96.14	94.55	10.28	97.92	86.89

Table 5. OOD detection experiments settings and results on plant village dataset. The ID dataset includes healthy leaves from 12 species. The six OOD datasets contain diseased leaves from 11 species. The results are based on the average of the six OOD datasets. Bold indicates the best performance.

OOD detection settings

Unlike the OSR setting, OOD detection takes one dataset as the ID and identifies several other datasets as OOD, with a requirement that there should be no overlapping classes between the ID and OOD datasets. However, despite the different dataset configurations for the two subtasks, both approaches essentially tackle the same challenge of detecting semantic shifts. As Plant Village contains disease types from 14 species, with 12 having healthy leaf samples, it is convenient to set up the OOD task. We designated the healthy leaf samples from the 12 categories as the ID dataset, while the remaining disease classes were grouped by species to form six OOD datasets. Table 5 shows the data split and experimental results.

The visual prompt method demonstrated stability across different OOD detection methods. For example, the AUROC exhibited a stable distribution ranging from 97.60% to 98.25%. With the stability of visual prompts, practitioners can confidently select evaluation methods, as these stable cues offer reliable and consistent results regardless of the circumstances, thereby enhancing the reliability and consistency of evaluations. Furthermore, OOD detection methods based on entropy, variance, and maximum softmax probability achieved better FPR@95 in the fully fine-tuning paradigm, while showing a significant gap in AUPR compared to the visual prompt-based method. This suggests that fully fine-tuning may lead to superior performance under specific threshold conditions. However, by visualizing the uncertainty distribution, we argue that finding an appropriate threshold for methods such as those based on entropy is challenging. We will elaborate on this in the discussion section.

Few-shot OOD detection settings

Collecting plant disease data presents numerous challenges. The complexity and diversity of plant diseases, driven by various pathogens and environmental factors, require extensive time and resources for accurate identification and data collection. Additionally, the lack of specialized disease recognition knowledge among farmers and the general public poses a barrier to data collection. Further compounding these challenges is the decentralized nature of data sources, which typically include farmers, agricultural institutions, and research organizations, complicating data integration and analysis. In practical applications, the availability of labeled data is typically limited. Few-shot learning addresses this limitation by enabling models to learn effectively from a small number of samples. This approach is particularly valuable for developing accurate models in scenarios with constrained data availability. Therefore, assessing the performance of out-of-distribution (OOD) detection under few-shot learning conditions is vital for evaluating the safety and robustness of machine learning models in this context.

Our study investigated the OOD detection performance of three fine-tuning paradigms-fully fine-tuning, linear probe tuning, and visual prompt fine-tuning-in a few-shot learning environment. We tested these paradigms using 2, 4, 8, and 16 shots to evaluate their effectiveness. Figure 6 illustrates the performance trends of these three fine-tuning methods across different shot settings. The visual prompt tuning paradigm consistently outperformed the other methods in threshold-free evaluation metrics, such as AUROC, AUPR, and ID accuracy. Notably, in the 2, 4, and 8-shot scenarios, visual prompt tuning demonstrated an advantage in all evaluation metrics. These results underscore that selecting the appropriate fine-tuning paradigm can significantly enhance the effectiveness of existing OOD detection methods. For example, when the dataset scale is small, using prompt tuning can significantly improve ID accuracy and OOD detection performance.

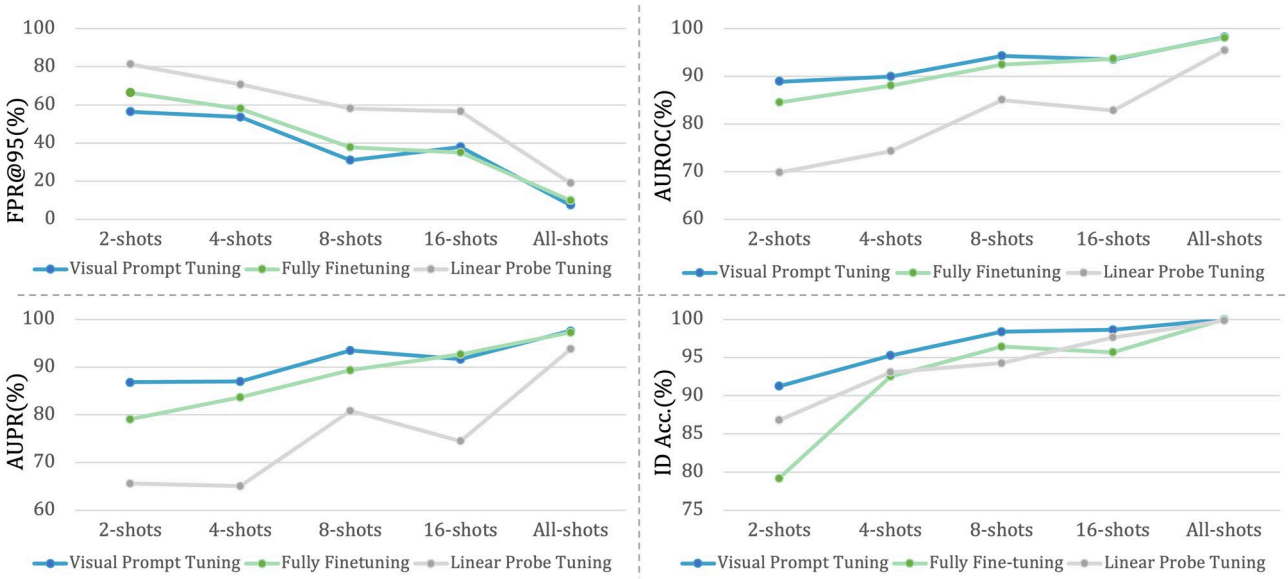


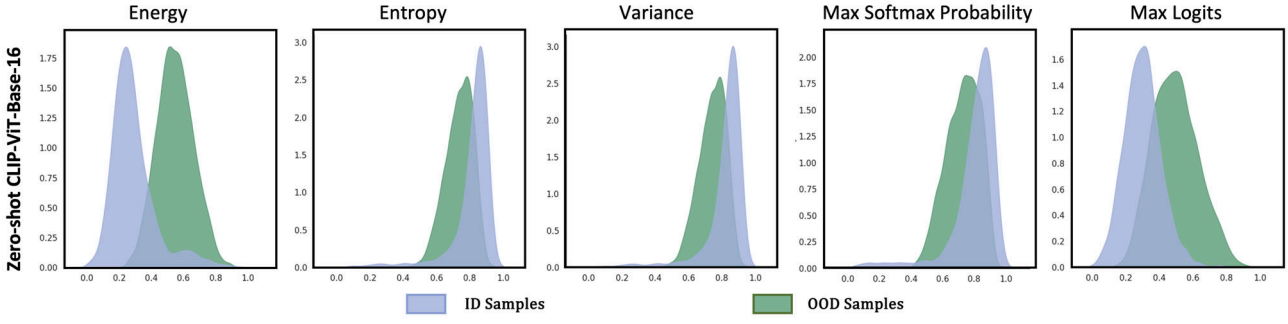
Figure 6. Performance on few-shot setting. To avoid confusion, the default OOD detection method is based on the maximum logits. VPT consistently leads across various evaluation metrics and experimental settings, especially under few-shot conditions.

Discussion

This section discusses the limitations of the current state-of-the-art OOD detection methods based on visual language models (VLMs) in the context of unrecognized plant diseases. Additionally, we explore the issue of threshold selection in different OOD detection methods through qualitative analysis.

Zero-shot VLM and other lightweight fine-tuning approaches have shown promise in improving OOD performance in natural language processing, as evidenced by recent research[13]. Specifically, visual language models leveraging high-quality pre-trained features have demonstrated robustness, particularly in scenarios involving significant distribution shifts. For instance, on the PETS37 dataset[42], using category names as language prompts sufficed to differentiate between ID and OOD data. However, the challenge of creating effective language prompts for plant disease data is markedly more complex. Our testing of three language prompts on the CLIP-ViT-base-16 and CLIP-ViT-large-14 models[19] yielded AUROC scores between 50% and 60%, suggesting that the model's performance was no better than random guessing (Table 1). Additionally, we assessed the vision pre-trained model using various OOD detection methods, with unexpected results: methods based on maximum logits and energy outperformed the MSP, previously considered the best approach in the MCM[13]. To visually display this difference, we visualized the distribution of uncertainty scores in Fig. 7. Compared to MSP, the uncertainty scores based on max-logits show a clear separation.

We acknowledge that comparing zero-shot visual language pre-trained models with fine-tuned visual models may not be fair. By efficiently fine-tuning these visual language models, OOD detection performance can be significantly improved. Therefore, we have reimplemented these methods and compared their performance in unknown plant disease recognition. Table 6 and Fig. 8 present the experimental results of the relevant methods. The results indicate that even when employing text prompt methods such as context optimization (CoOp)[20] and conditional context optimization (CoOpOp)[21] to generate adaptive text prompts for plant diseases automatically,



Note: In-distribution Dataset include 12 classes Healthy Leaves, and out-of-distribution dataset is 3 grape diseases introduced in Table 5.

Figure 7. Kernel density estimation plots of zero-shot out-of-distribution detection models on plant disease datasets. The language prompt type is:"photo of a Class names". Uncertainty scores were calculated using MCM.

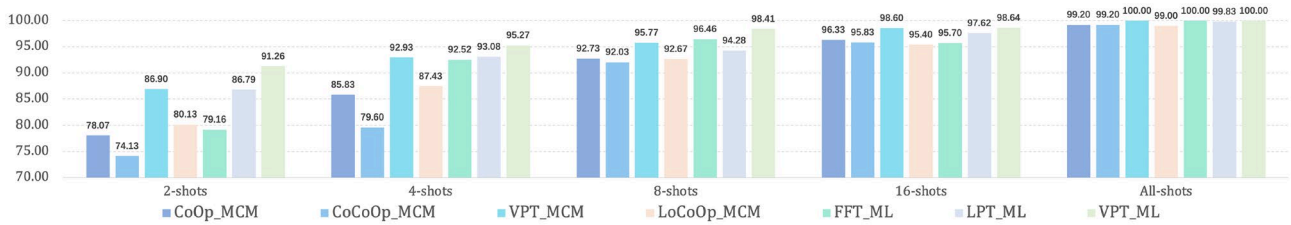


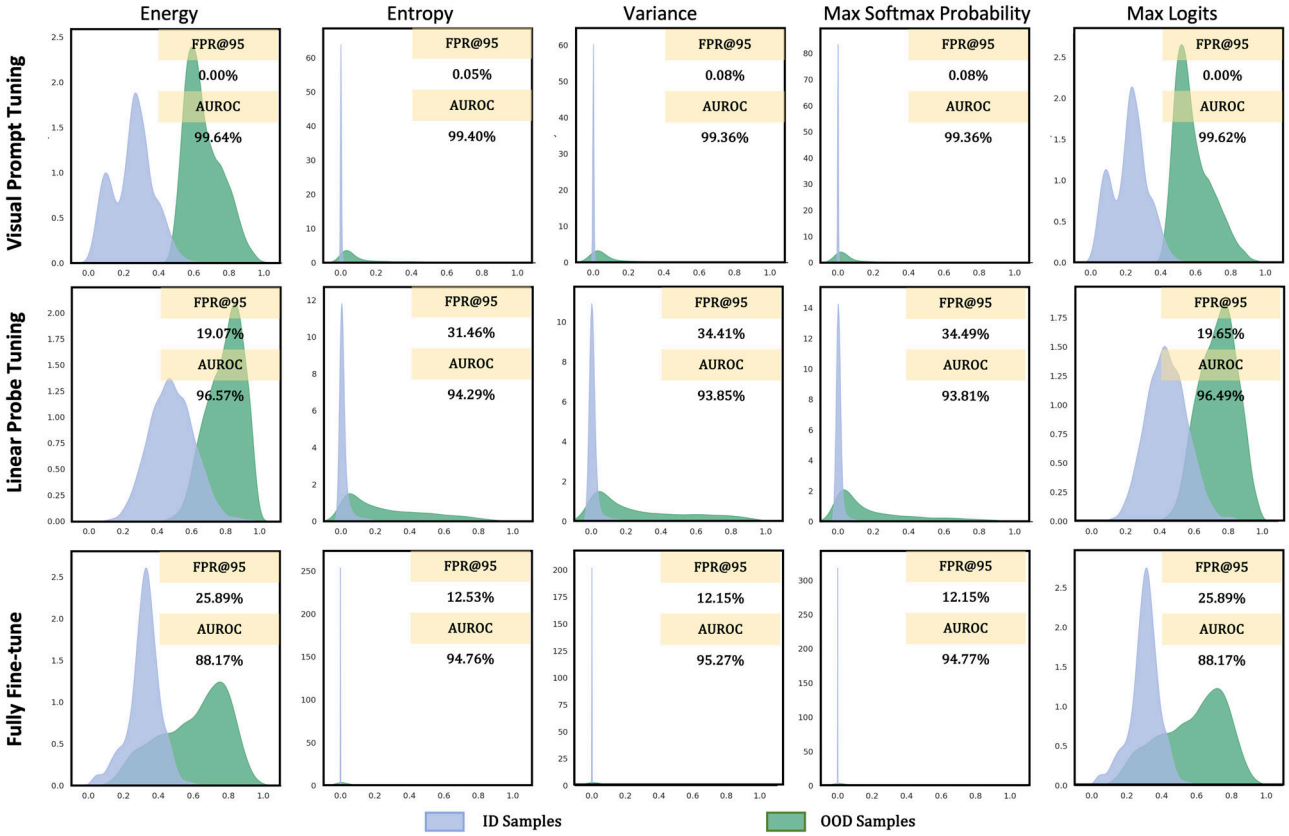
Figure 8. Comparison with CLIP-based vision-language models on ID accuracy. Even with fine-tuning through text prompting methods, CLIP still fails to achieve high accuracy, which may lead to failure in unknown plant disease recognition.

Method	Context prompt	FPR@95↓ / AUROC↑ at different shots				
CLIP-Based		2 -shot	4-shot	8-shot	16-shot	All-shot
$CoOp_{MCM}$[20]	Learnable Prompt + [CLS]	93.50/54.40	86.93/56.91	89.32/60.09	81.94/66.87	68.97/75.28
$CoCoOp_{MCM}$[21]	Learnable Prompt + [CLS]	91.81/60.92	85.39/63.99	77.37/76.63	64.89/81.20	40.08/88.61
VPT_{MCM}[30]	a photo of a + [CLS]	85.01/68.24	72.90/73.66	61.74/84.61	51.55/84.20	15.67/96.23
$LoCoOp_{MCM}$[15]	Learnable Prompt + [CLS]	90.33/61.46	86.16/65.01	82.32/70.55	78.54/71.31	68.19/76.88
ImageNet-21k-Based		2-shot	4-shot	8-shot	16-shot	All-shot
$FFT_{MaxLogits}$	–	66.42/84.49	57.90/88.04	37.68/92.43	**35.07/93.66**	10.28/97.92
$LPT_{MaxLogits}$	–	81.33/69.79	70.72/74.28	57.99/85.00	56.62/82.83	15.09/96.14
$VPT_{MaxLogits}$	–	**56.42**/88.85	**53.67/89.90**	**30.98/94.30**	37.97/93.52	**7.40/98.25**

Table 6. Comparison with CLIP-based vision-language models on FPR@95 and AUROC scores. [CLS] denotes the class name, where [CLS] denotes class names. Bold indicates the best performance.

visual language models consistently underperform. Notably, when we deploy visual prompt fine-tuning in the visual language models, there is a notable improvement in out-of-distribution detection performance. Nonetheless, this performance is still far below that of purely visual pre-trained models. Moreover, visual language pre-trained models based on CLIP also perform worse than purely visual pre-trained models on PlantVillage in ID accuracy, as shown in Fig. 8. We believe this may be due to the inadequacies of text prompts to generate optimal prompts for plant diseases. We will study this issue in our future work.

Our analysis also extended to the effectiveness of different OOD detection methods in distinguishing between ID and OOD data. Fig. 9 illustrates the separation results of five post-hoc methods under three fine-tuning paradigms. We noted that for methods based on entropy, variance, and MSP, uncertainty scores for both ID and OOD data tended to cluster around 0, indicating high confidence in classifying all samples. In fully fine-tuned models, OOD detection methods based on entropy, variance, and MSP performed better than those based on energy and maximum logits (Table 5). However, identifying an appropriate threshold for these three methods is challenging, as minor threshold variations can lead to an inability to distinguish ID from OOD data. Conversely, selecting suitable thresholds for energy-based and maximum logits-based methods is comparatively more straightforward. We provide example analyses in Fig. 10 to visually demonstrate these challenges. We use 0.5 as the threshold. If MSP is employed as the OOD detection method, the model classifies all images as known. However, using Max Logits significantly ameliorates this issue. When using 0.05 as a threshold to separate known and unknown samples, methods based on MSP can achieve their full potential. However, such a threshold might be difficult to adapt to most datasets. Overall, although the method based on maximum logits is slightly inferior to the



Note: The ID dataset includes 12 classes of healthy leaves, and the OOD dataset includes 3 grape diseases introduced in Table 5.

Figure 9. Kernel density estimation plots for the three fine-tuning methods. Uncertainty scores were calculated by different methods. Evaluation metrics are added to the top right corner of each subplot.

MSP method in performance metrics, we suggest choosing OOD detection methods based on logits or energy wherever possible, due to the ease of setting a universal threshold across different datasets.

In our few-shot OOD detection experiments, the visual prompt-based method remarkably achieved over 90% ID recognition accuracy with just two samples per category. In the 2, 4, and 8-shot setup, visual prompts significantly surpassed linear probe tuning and fully fine-tuning methods in all performance metrics, which indicates that the choice of fine-tuning method significantly impacts OOD detection performance. Overall, the method based on visual prompts is more effective in small-sample OOD detection tasks because it retains more of the pre-trained model's general knowledge and learns domain-specific knowledge of the downstream task. Our research enhances plant disease recognition systems' safety, reliability, and performance by providing insights into handling unknown diseases and selecting appropriate fine-tuning paradigms in scenarios with limited data and uncertainty.

Conclusion

Identifying and rejecting disease categories not encountered during the training phase are crucial for the reliability of systems and applications. This paper establishes benchmarks for open-set recognition (OSR), out-of-distribution (OOD) detection, and few-shot OOD detection, all related to the task of recognizing unknown classes, using five plant disease datasets. We conducted comprehensive benchmark testing on five OOD detection methods and three fine-tuning paradigms. Our extensive experiments have demonstrated the visual prompt method as the most effective approach for recognizing unknown diseases, particularly in few-shot OOD detection scenarios. By studying the impact of fine-tuning paradigms on the task of detecting unknown plant diseases, we argue that choosing the appropriate fine-tuning method can directly enhance the performance of OOD detection methods while avoiding additional computational costs. We hope that future researchers will test the performance of their novel OOD detection methods under various fine-tuning paradigms, potentially leading to unexpected performance improvements in fine-grained datasets, such as those involving plant diseases.

Despite these achievements, there are still some limitations in our current work. For instance, the challenges posed by visual language models in recognizing unknown plant diseases have not been fully addressed. We will focus on this issue in our future work. Additionally, we have not yet explored the OOD detection performance of other fine-tuning methods within plant disease recognition tasks, such as multilayer perceptron, bias[43], and partial fine-tuning[44]. We encourage other researchers to investigate these methods, potentially contributing further to the advancement of robust and efficient plant disease detection and classification systems.

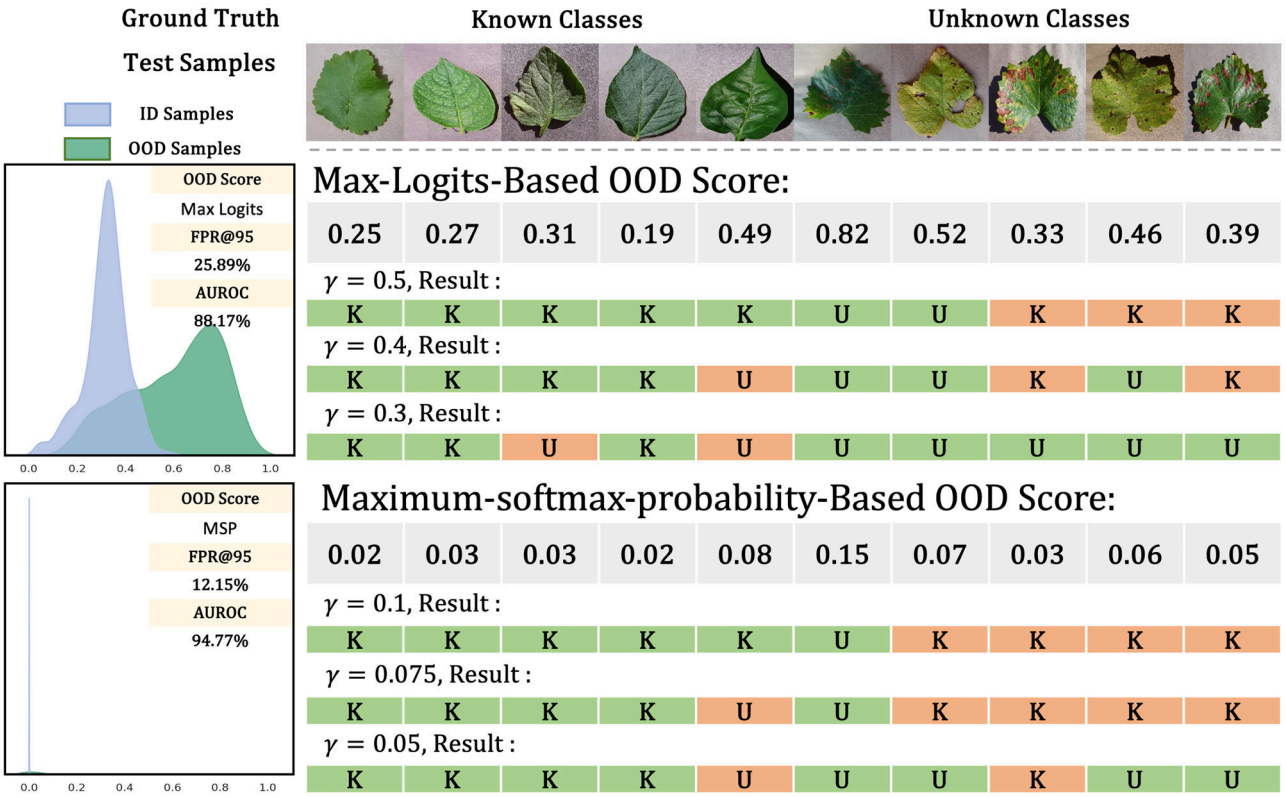


Figure 10. Qualitative analysis. We compared the decision of MSP-based OOD score and Max logits-based OOD score with gray background. The experimental setup involved using the fully fine-tuning strategy on the plant village dataset. 'K' and 'U' represent known and unknown respectively. A green background indicates correct predictions, while orange denotes incorrect predictions.

Data availability and access

The datasets we used are all public datasets, and the publishers encourage researchers to use their datasets for academic research. We have ensured that all the datasets used are easily accessible and have provided detailed information for their retrieval. The datasets employed in our study are available at the following locations: **Cotton Disease Dataset:** This dataset is available on Kaggle and can be accessed through the link https://www.kaggle.com/datasets/dhamur/cotton-plant-disease. Mango Disease Dataset: The dataset pertaining to mango diseases is also hosted on Kaggle and can be reached via https://www.kaggle.com/datasets/aryashah2k/mango-leaf-disease-dataset. Strawberry Disease Dataset: For research involving strawberry diseases, the dataset is available at https://www.kaggle.com/datasets/usmanafzaal/strawberry-disease-detection-dataset. Tomato Disease Dataset and Plant Village Dataset: These datasets can be found at the same location, accessible through the link https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw.

Received: 9 January 2024; Accepted: 5 July 2024
Published online: 02 August 2024

References

1. Carroll, C. L., Carter, C. A., Goodhue, R. E. & Lawell, C.-Y. Crop disease and agricultural productivity: Evidence from a dynamic structural model of verticillium wilt management. In *Agricultural Productivity and Producer Behavior*, 217–249 (University of Chicago Press, 2018).
2. Savary, S. *et al.* The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* **3**, 430–439 (2019).
3. Li, L., Zhang, S. & Wang, B. Plant disease detection and classification by deep learning-a review. *IEEE Access* **9**, 56683–56698 (2021).
4. Shafik, W., Tufail, A., Namoun, A., De Silva, L. C. & Apong, R. A. A. H. M. A systematic literature review on plant disease detection: Techniques, dataset availability, challenges, future trends, and motivations. *IEEE Access* **11**, 59174–59203 (2023).
5. Nazki, H., Yoon, S., Fuentes, A. & Park, D. S. Unsupervised image translation using adversarial networks for improved plant disease recognition. *Comput. Electron. Agric.* **168**, 105117 (2020).
6. Tian, L. *et al.* VMF-SSD: A novel v-space based multi-scale feature fusion SSD for apple leaf disease detection. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 2016–2028 (2022).
7. Dong, J. *et al.* Data-centric annotation analysis for plant disease detection: Strategy, consistency, and performance. *Front. Plant Sci.* **13**, 1037655 (2022).
8. Dong, J., Fuentes, A., Yoon, S., Kim, H. & Park, D. S. An iterative noisy annotation correction model for robust plant disease detection. *Front. Plant Sci.* **14**, 1238722 (2023).
9. Du, X., Wang, Z., Cai, M. & Li, Y. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. *International Conference on Learning Representations* (ICLR, 2022).
10. Xiong, H. *et al.* From open set to closed set: Supervised spatial divide-and-conquer for object counting. *Int. J. Comput. Vis.* **131**, 1722–1740 (2023).

11. Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016).
12. Fuentes, A., Yoon, S., Kim, T. & Park, D. S. Open set self and across domain adaptation for tomato disease recognition with deep learning techniques. *Front. Plant Sci.* **12**, 758027 (2021).
13. Ming, Y. *et al.* Delving into out-of-distribution detection with vision-language representations. *Adv. Neural Inf. Process. Syst.* **35**, 35087–35102 (2022).
14. Ming, Y. & Li, Y. How does fine-tuning impact out-of-distribution detection for vision-language models?. *Int. J. Comput. Vis.* **132**(2), 596–609 (2024).
15. Miyai, A., Yu, Q., Irie, G. & Aizawa, K. LoCoOp: Few-shot out-of-distribution detection via prompt learning. In *Thirty-Seventh Conference on Neural Information Processing Systems* (2023).
16. Fort, S., Ren, J. & Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *Adv. Neural Inf. Process. Syst.* **34**, 7068–7081 (2021).
17. Huang, R. & Li, Y. MOS: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8710–8719 (2021).
18. Lee, K., Lee, K., Lee, H. & Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, Vol. 31 (2018).
19. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 8748–8763 (PMLR, 2021).
20. Zhou, K., Yang, J., Loy, C. C. & Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **130**, 2337–2348 (2022).
21. Zhou, K., Yang, J., Loy, C. C. & Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825 (2022).
22. Yang, J., Zhou, K., Li, Y. & Liu, Z. Generalized out-of-distribution detection: A survey. *Int. J. Comput. Vis.* 1–28 (2024).
23. Liang, S., Li, Y. & Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations* (2018).
24. Hendrycks, D. *et al.* Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 8759–8773 (PMLR, 2022).
25. Liu, W., Wang, X., Owens, J. & Li, Y. Energy-based out-of-distribution detection. *Adv. Neural. Inf. Process. Syst.* **33**, 21464–21475 (2020).
26. Lin, Z., Roy, S. D. & Li, Y. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 15313–15323 (2021).
27. Hendrycks, D., Lee, K. & Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *international Conference on Machine Learning*, 2712–2721 (PMLR, 2019).
28. Kirillov, A. *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026 (ICCV, 2023).
29. Kornblith, S., Shlens, J. & Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2661–2671 (2019).
30. Jia, M. *et al.* Visual prompt tuning. In *European Conference on Computer Vision*, 709–727 (Springer, 2022).
31. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2020).
32. Dhamodharan. Cotton plant disease (2023).
33. Ahmed, S. I. *et al.* MangoLeafBD: A comprehensive image dataset to classify diseased and healthy mango leaves. *Data Brief* **47**, 108941 (2023).
34. Afzaal, U., Bhattarai, B., Pandeya, Y. R. & Lee, J. An instance segmentation model for strawberry diseases based on mask R-CNN. *Sensors* **21**, 6565 (2021).
35. 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).
36. Chen, Z. *et al.* Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations* (ICLR, 2023).
37. Yao, Y. *et al.* W-transformer: Accurate cobb angles estimation by using a transformer-based hybrid structure. *Med. Phys.* **49**, 3246–3262 (2022).
38. Ryu, S., Koo, S., Yu, H. & Lee, G. G. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 714–718 (2018).
39. Powers, D. M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020).
40. Gunawardana, A. & Shani, G. A survey of accuracy evaluation metrics of recommendation tasks. *J. Mach. Learn. Res.* **10**, 2935–2962 (2009).
41. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, 2009).
42. Parkhi, O. M., Vedaldi, A., Zisserman, A. & Jawahar, C. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3498–3505 (IEEE, 2012).
43. Zaken, E. B., Ravfogel, S. & Goldberg, Y. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199 (2021).
44. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, Vol. 27 (2014).

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2019R1A6A1A09031717); by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (RS-2024-00360581); and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (NRF-2021R1A2C1012174).

Author contributions

J.D.: Conceptualization, Methodology, Software, Writing-Original draft preparation. A.F.: Writing- Reviewing and Editing, H.Z.: Software and Methodology. Y.J.: Conceptualization, Methodology. S.Y.: Supervision, Conceptualization, Methodology. DS.P.: Conceptualization, Methodology.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.Y. or D.S.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.