

Benchmarking In-the-Wild Multimodal Plant Disease Recognition and A Versatile Baseline

Anonymous Authors

APPENDIX

1 IMAGE SAMPLES

We present exemplars of images in PlantVillage [4], PlantDoc [8] and our curated PlantWild dataset, as shown in Figure 1. PlantVillage is composed of laboratory images with controlled background and lighting conditions. PlantDoc consists of plant disease images captured from in-the-wild environments, but it falls short in scale. In contrast, PlantWild contains in-the-wild images and includes a significantly broader range of classes compared to existing plant disease datasets.

2 DETAILED STATISTICS OF PLANTWILD

PlantWild consists of 18,542 images across 89 classes. The class with the most images includes 589 images and the class with the fewest images includes 44 images. The detailed statistics of PlantWild are shown in Table 2, including all class names and the number of images in each class.

3 DESCRIPTIVE PROMPTS GENERATION

Figure 2 illustrates the process of prompt generation. To enhance the diversity of generated descriptive prompts for the text encoder

of CLIP [6], we design different commands as the input to GPT-3.5 [2]. The commands are presented as follows:

- Summarize visual characteristics of [CLASS] with less than [LENGTH] words.
- How to identify [CLASS]? Answer within [LENGTH] words.
- Use less than [LENGTH] words to outline a photo of [CLASS].

[CLASS] refers to the names of corresponding classes, while [LENGTH] stands for the maximum allowable number of words for prompt generation. If [CLASS] is a disease class, we leverage two additional commands to obtain descriptions of symptoms caused by the disease in early and late stages:

- Describe the early symptoms of [CLASS] using no more than [LENGTH] words.
- Describe the late symptoms of [CLASS] using no more than [LENGTH] words.

GPT-3.5 usually generates excessively long prompts to describe the characteristics of plant diseases in detail. CLIP may encounter difficulties in extracting useful information from excessively long sentences, leading to a decline in classification performance. Therefore, it is necessary to limit the length of generated prompts. In particular, we set the length limitation [LENGTH] at 25 words.

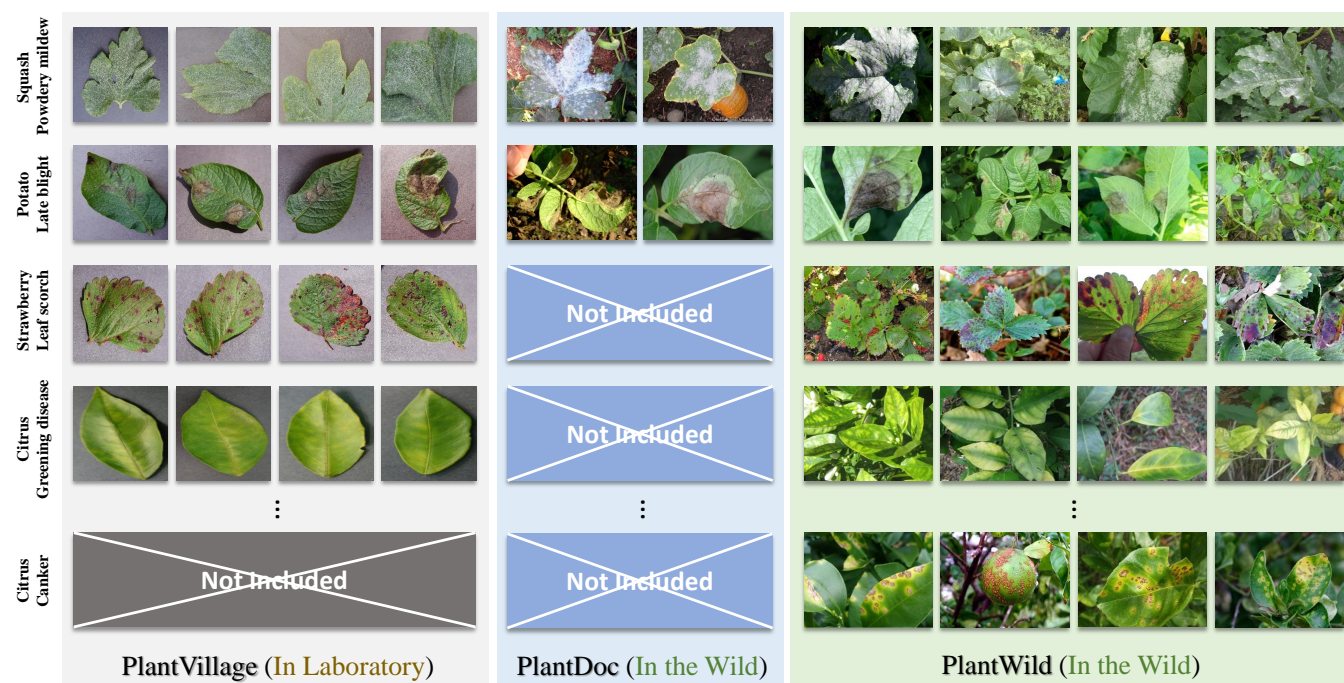


Figure 1: Comparisons of our curated dataset PlantWild and existing PlantVillage and PlantDoc. PlantWild consists of in-the-wild images and contains the largest number of disease classes.

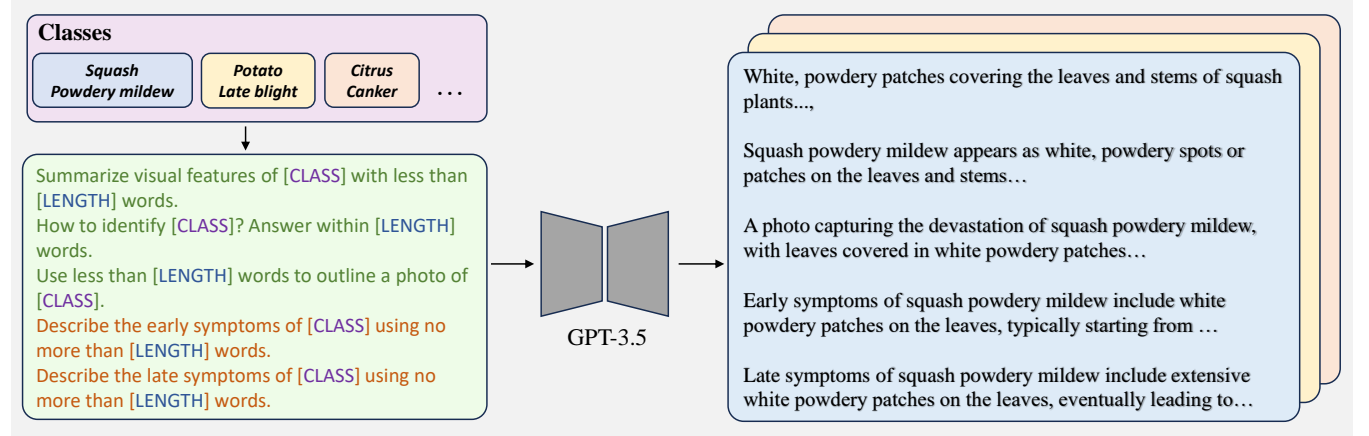


Figure 2: The generation process of textual prompts for each class in PlantWild. For healthy classes, we leverage three commands as inputs to GPT-3.5 for generation. For disease classes, we utilize two additional commands to obtain symptom descriptions at different periods. We set a length limitation as 25 for generated prompts.

Methods	<i>PlantVillage</i>				<i>PlantDoc</i>				<i>PlantWild</i>			
	Acc	M-P	M-R	M-F1	Acc	M-P	M-R	M-F1	Acc	M-P	M-R	M-F1
AgglomerativeClustering [11]	97.66	97.18	96.55	96.77	68.51	67.29	67.13	66.34	66.41	64.63	62.90	62.92
SpectralClustering [7]	97.70	97.34	96.42	96.81	68.32	67.16	66.90	66.34	66.44	64.09	62.78	63.05
Bi K-means [9]	97.44	97.13	96.16	96.51	69.31	67.92	67.98	66.47	66.58	63.15	62.04	61.80
BIRCH [12]	97.17	96.80	95.69	96.02	69.70	69.31	68.46	66.47	66.79	64.03	63.12	63.05
OPTICS [1]	97.14	96.89	95.08	95.70	68.71	67.45	67.72	67.15	66.82	64.01	63.35	63.24
MeanShift [3]	97.29	96.91	95.68	96.16	68.91	67.93	97.75	66.27	66.88	64.03	63.12	63.05
K-means [5]	97.72	97.40	96.44	96.83	69.90	69.92	68.97	68.87	67.20	64.03	62.64	62.84

Table 1: Classification results of our proposed baseline MVPDR with various grouping techniques. Our baseline demonstrates strong robustness across different grouping techniques and achieves the best overall performance with K-means.

4 FALIURE CASE ANALYSIS

We investigate the failure cases of our proposed MVPDR where images are misclassified. Figure 3 illustrates some representative failure cases. Failure case 1 in Figure 3 (Left) illustrates that mild symptoms of cucumber powdery mildew may not be prominent enough to clearly distinguish it from healthy cucumber leaf, resulting in misclassification. Different diseases with similar characteristic presentations can also raise difficulties for classification, see failure case 2 in Figure 3 (Middle). In addition, according to Figure 3 (Right), for citrus canker that can occur on leaves and fruit, we observe that all images of diseased fruit are correctly identified while some diseased leaves are misclassified.

5 CONFUSION MATRIXES OF MVPDR AND DHBP

To get further insights from the classification results, we generate confusion matrixes based on the results of MVPDR on PlantVillage, PlantDoc and PlantWild. According to the results in Figure 4, PlantWild is more challenging than other datasets for classification tasks. In addition, we also produce confusion matrixes for the second best-performing competing method DHBP [10], as shown in Figure 5. Comparing Figure 4 and Figure 5, it can be observed that

DHBP has superiority on PlantVillage while MVPDR shows better performance on PlantDoc and PlantWild datasets.

6 MVPDR WITH DIFFERENT GROUPING TECHNIQUES

We explore the influence of different grouping techniques in our baseline MVPDR. We utilize K-means [5], Bi K-means [9], MeanShift [3], BIRCH [12], Agglomerative Clustering [11], Spectral Clustering [7] and OPTICS [1] for visual prototype construction, then evaluate the classification performance.

Specifically, for partition-based clustering methods such as K-means and Spectral Clustering that need to know how many clusters the data should be divided into, we follow the main experiment setups and specify the number of clusters to 16. In contrast, hierarchical clustering and density-based clustering methods such as MeanShift and OPTICS do not need to specify the number of clusters beforehand, as these methods can automatically determine the number of clusters based on data point densities or distances. According to the results in Table 1, MVPDR is robust across different grouping techniques. By leveraging K-Means, MVPDR achieves the highest accuracy on PlantWild and the best overall performance on all three datasets.

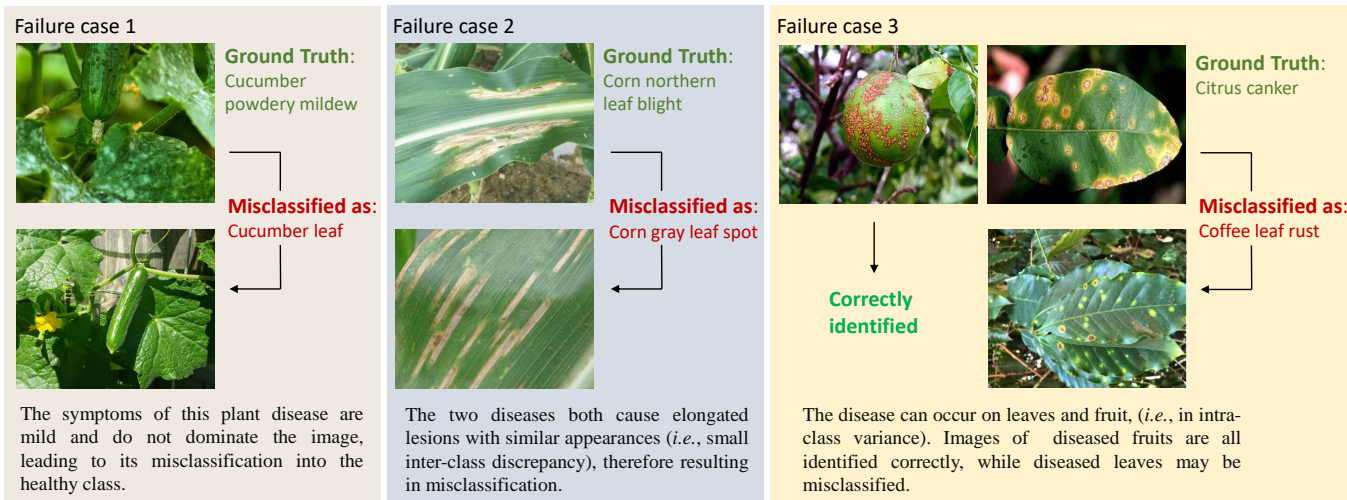


Figure 3: Illustrations of some representative failure cases of our proposed baseline.

REFERENCES

[1] Mihael Ankerst, M Breunig, Hans-Peter Kriegel, R Ng, and J Sander. 2008. Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD*.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Keinosuke Fukunaga and Larry Hostetler. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory* (1975), 32–40.

[4] David Hughes, Marcel Salathé, et al. 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060* (2015).

[5] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 281–297.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. 8748–8763.

[7] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* (2000), 888–905.

[8] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. 2020. PlantDoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 249–253.

[9] Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. (2000).

[10] Dongfang Wang, Jun Wang, Zhuang Ren, and Wenrui Li. 2022. DHBP: A dual-stream hierarchical bilinear pooling model for plant disease multi-task classification. *Computers and Electronics in Agriculture* (2022), 106788.

[11] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* (1963), 236–244.

[12] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: an efficient data clustering method for very large databases. *ACM sigmod record* (1996), 103–114.

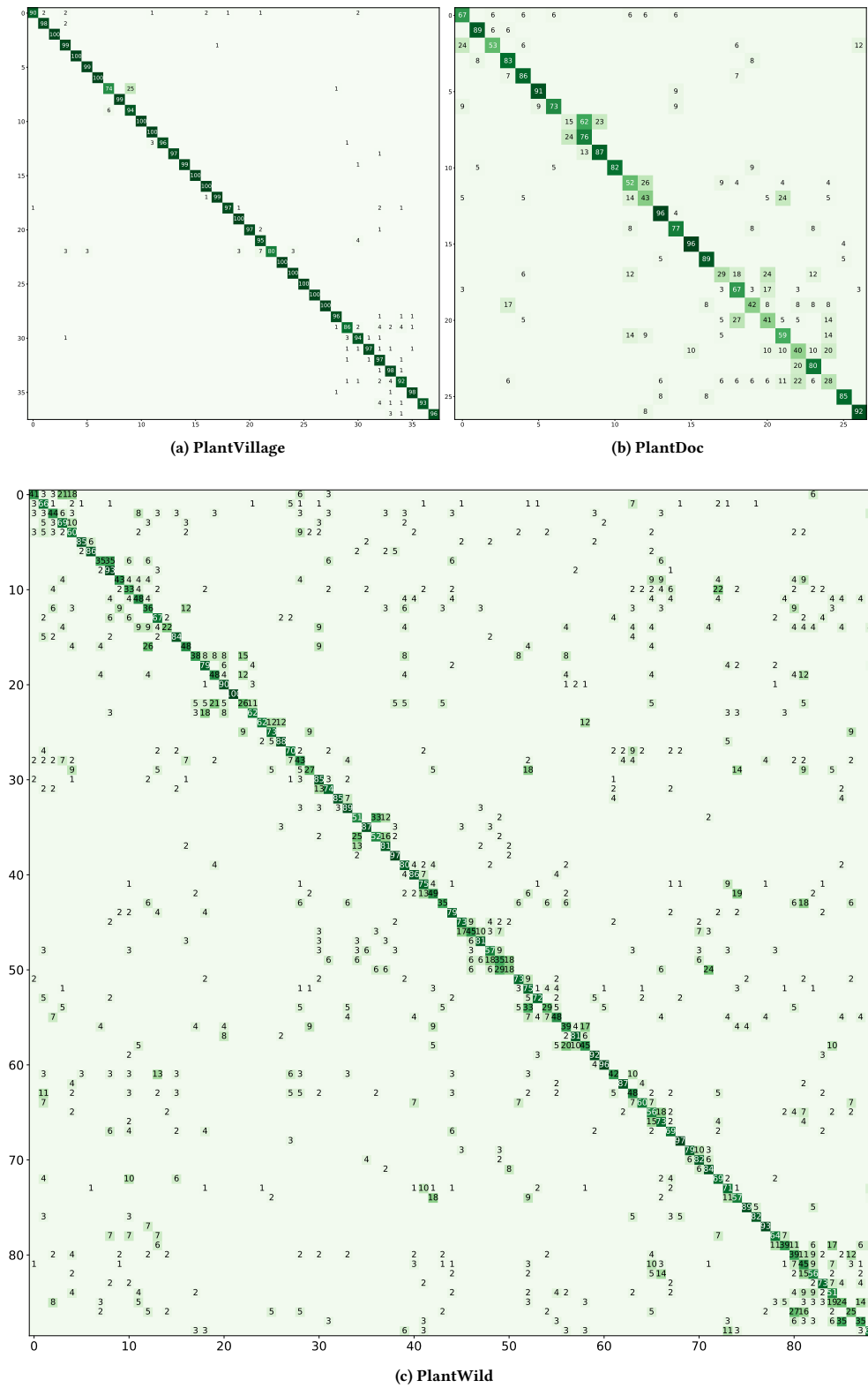


Figure 4: Confusion matrix results of MVPDR on PlantVillage, PlantDoc and PlantWild. The value in each grid represents a percentage. It illustrates that PlantWild is more challenging than other datasets for classification tasks.

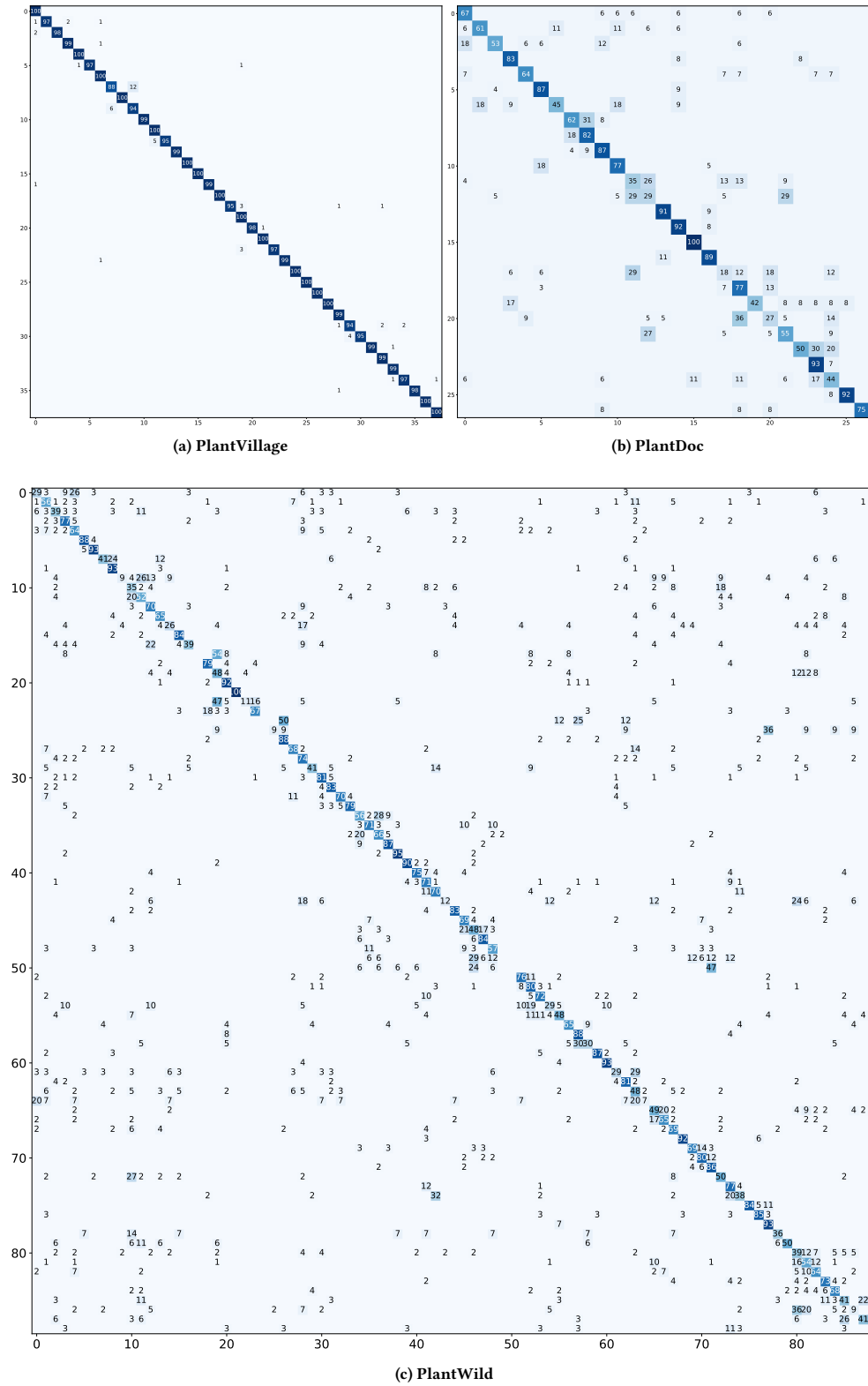


Figure 5: Confusion matrix results of DHBP [10] on PlantVillage, PlantDoc and PlantWild. The value in each grid represents a percentage.

Class name	#Images	Class name	#Images	Class name	#Images
1.apple black rot	173	31.citrus canker	535	61.maple tar spot	139
2.apple leaf	444	32.citrus greening disease	237	62.peach leaf	157
3.apple mosaic virus	181	33.coffee leaf	138	63.peach leaf curl	235
4.apple rust	308	34.coffee leaf rust	190	64.plum leaf	325
5.apple scab	292	35.corn gray leaf spot	216	65.plum pocket disease	76
6.banana leaf	243	36.corn leaf	156	66.potato early blight	227
7.banana panama disease	216	37.corn northern leaf blight	224	67.potato late blight	240
8.basil downy mildew	86	38.corn rust	237	68.potato leaf	249
9.basil leaf	589	39.corn smut	293	69.raspberry leaf	180
10.bean halo blight	115	40.cucumber angular leaf spot	251	70.rice blast	148
11.bean leaf	258	41.cucumber bacterial wilt	143	71.rice leaf	252
12.bean mosaic virus	125	42.cucumber leaf	348	72.rice sheath blight	250
13.bean rust	165	43.cucumber powdery mildew	236	73.soybean leaf	242
14.bell pepper leaf	240	44.eggplant cercospora leaf spot	88	74.squash leaf	410
15.bell pepper bacteria leaf spot	117	45.eggplant leaf	240	75.squash powdery mildew	281
16.blueberry leaf	281	46.garlic leaf	228	76.strawberry anthracnose	98
17.blueberry rust	117	47.garlic leaf blight	147	77.strawberry leaf	199
18.broccoli downy mildew	65	48.garlic rust	160	78.strawberry leaf scorch	76
19.broccoli leaf	269	49.ginger leaf	177	79.tobacco leaf	71
20.cabbage alternaria leaf spot	128	50.ginger leaf spot	87	80.tobacco mosaic virus	91
21.cabbage leaf	464	51.ginger sheath blight	87	81.tomato bacterial leaf spot	280
22.carrot cavity spot	74	52.grape black rot	229	82.tomato early blight	346
23.cauliflower alternaria leaf spot	98	53.grape downy mildew	395	83.tomato late blight	295
24.cauliflower leaf	195	54.grape leaf	201	84.tomato leaf	226
25.celery anthracnose	44	55.grape leaf spot	106	85.tomato leaf mold	239
26.celery early blight	55	56.grapevine leafroll disease	138	86.tomato mosaic virus	189
27.celery leaf	212	57.lettuce downy mildew	118	87.tomato septoria leaf spot	220
28.cherry leaf	286	58.lettuce leaf	244	88.tomato yellow leaf curl virus	171
29.cherry leaf spot	230	59.lettuce mosaic virus	100	89.zucchini yellow mosaic virus	181
30.cherry powdery mildew	114	60.maple leaf	316	Total	18542

Table 2: The number of images in each class of PlantWild. Our PlantWild dataset contains a significantly larger number of classes compared to PlantVillage and PlantDoc.