Validation of Prerequisites for Correct Performance Evaluation of Image-based Plant Disease Diagnosis using Reliable 221K Images Collected from Actual Fields

Shogo Shibuya [1], Quan Huu Cap [1], Shunta Nagasawa [1],
Satoshi Kagiwada [2], Hiroyuki Uga [3], Hitoshi Iyatomi [1]
[1]Applied Informatics, Graduate School of Science and Engineering, Hosei University, Tokyo, Japan
[2]Clinical Plant Science, Faculty of Bioscience and Applied Chemistry, Hosei University, Tokyo, Japan
[3]Saitama Agricultural Technology Research Center, Saitama, Japan
{syogo.shibuya.5u@stu., huuquan.cap.75@, shunta.nagasawa.2u@stu., kagiwada@, iyatomi@}hosei.ac.jp
uga.hiroyuki@pref.saitama.lg.jp

Abstract

Although many image-based plant disease diagnosis systems have reported high diagnostic performance recently, most of them do not seem to have a proper separation between the training and evaluation images. Because of the potential similarity of images taken in the same field, the true performance of a system where the same field is used training and evaluation images is much worse than it appears. However, no systematic evaluation based on large-scale data has been conducted so far. To suppress overfitting due to such similarity, several attempts have been made to detect regions of interest (ROI), such as leaves, in advance, but no systematic studies have been conducted on their effectiveness. In this study, we used a total of 221,842 leaf images of four crops from 24 prefectures with reliable labels to investigate (i) the performance bias due to evaluation within the same farm and (ii) the effect of the ROI detection on the performance. As a result, even if a large number of training images with sufficient resolution are prepared, diagnostic performance for images in fields different from the training images is greatly degraded due to large differences in image characteristics, i.e., covariate shift. In this situation, the benefit of ROI detection became smaller.

1 Introduction

Pests and diseases are a threat to food security, with the United Nations Food and Agriculture Organization (FAO) estimating that they cause losses of 20-40% of the world's food production (FAO 2020). Thus, early detection of diseases and their appropriate treatment are essential to minimizing the damage they cause. Since plant disease diagnosis by experts is expensive, image-based automatic plant disease diagnosis techniques centered on convolutional neural networks (CNNs) have been proposed in recent years (Mohanty, Hughes, and Salathé 2016; Durmuş, Güneş, and Kırcı 2017; Wang, Sun, and Wang 2017; Brahimi, Boukhalfa, and Moussaoui 2017; Brahimi et al. 2018; Toda and Okura 2019; Fujita et al. 2016; Tani et al. 2018; Fuentes et al. 2017; Liu and Wang 2020; Yu and Son 2020; Ferentinos 2018; Suwa et al. 2019; Zekiwos, Bruck et al. 2021; Mithu et al. 2022).

Plant disease diagnosis is a fine-grained problem, where the disease symptoms are very small compared to the size of the image, and where differences in appearance between classes are slight although their variance within the class is often significant. Therefore, there are several challenges in image-based plant disease diagnosis. First, images taken at different locations (i.e., different farms) normally show significant dissimilarities in visual appearance, even for the same disease. This is due to different developing conditions, environments, and so on. Second, images of the same field, even for different diseases, are often very similar. Thus, plant disease identification in practical situations is a difficult task. As shown in Figure 1, (a) and (b) are the same disease (Bacterial Spot), but the appearance of the disease is different in other environments. On the other hand, (a) Bacterial Spot and (c) Downy Mildew taken in the same field are very similar, even though they are different diseases. For diseases with similar characteristics, such as Bacterial Spot and Downy Mildew, the difference in the environment may significantly impact identification results.

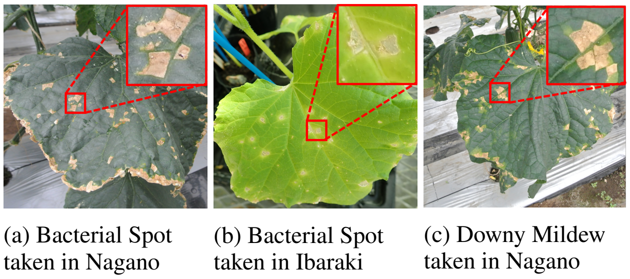

(a) Bacterial Spot taken in Nagano
(b) Bacterial Spot taken in Ibaraki
(c) Downy Mildew taken in Nagano

Figure 1: Cucumber leaf in in-farm environment. Even if (a) and (b) are of the same disease, some look different in other fields. On the other hand, some images, such as (a) and (c), of different diseases look very similar.

Despite the difficulties of handling practical in-farm images, many automated plant disease diagnosis studies have reported very high diagnostic accuracy (Fuentes et al. 2017; Liu and Wang 2020; Yu and Son 2020; Ferentinos 2018; Suwa et al. 2019; Zekiwos, Bruck et al. 2021; Mithu et al. 2022). From those studies, we strongly suspect that images were not properly separated into training and evaluation purpose. It is possible that images taken in the same field with similar composition, of the same subject, or the same subject taken at short intervals, were separated for training and evaluation. In other words, the inappropriate similarity between training and evaluation images may enhance the superficial performance. It has also been pointed out that diagnostic models with the same training and evaluation images achieve high apparent diagnostic accuracy, but reveal a clear performance degradation and overfitting when diagnosing diseases in other locations (Saikawa et al. 2019; Suwa et al. 2019; Boulent et al. 2019; Kanno et al. 2021; Cap et al. 2020, 2021). For example, Saikawa et al. (2019) revealed that a CNN model trained on nearly 36,000 images of cucumber leaves achieved an average accuracy of 97.5% on images from the same farm as the training data, but dropped to 40.3% on other farms. This suggests that it is important to carefully separate training and testing data sources for true performance evaluation, however, to the author's knowledge, this trend has only been reported for cucumber, and there are no reports on other crops. This is due to the difficulty of data collection, and until now, experiments using large amounts of data from various crops have not been feasible. This raises the question of whether or not this overfitting problem also happens with other crops.

In the meantime, some studies have claimed that background information of the image is the major cause of the overfitting problem (Saikawa et al. 2019; Ma et al. 2018; Zhang et al. 2021). They suggest that background similarity in a large region has a greater impact than disease-induced similarity, thus reducing diagnostic performance. They aimed for high identification performance for essentially unknown data, either by detecting foreground regions using background removal (Saikawa et al. 2019) or by detecting lesion sites (Zhang et al. 2021). Each was reported to have had some effect on disease diagnosis, but not enough to fill the large gap in the case of performance when the training and evaluation data were from the same farm and different farms. Moreover, research on the effect of background on disease diagnosis is very limited and has only been conducted on single crop thus far (i.e., cucumber). Hence, it cannot yet be concluded that background areas are the main reason for the overfitting problem. For stronger conclusions, we believe there is a need to study the effect of background of the images on a large amount of data from various crops. The goal that we should aim for in an automatic plant disease diagnosis system is a robust system that is stable even when the input data is taken in a different environment (different weather, background, shooting distance, etc.) from the training data. In this paper, we conduct two experiments on a large amount of disease leaf images from four crops (i.e., tomato, strawberry, cucumber, and eggplant) to answer the following questions:

- How important is it to use data taken in a different environment than the subdivision of the training data for performance evaluation?
- How much does the pre-detection of regions of interest (ROI) including symptoms of diseases affect the performance?

	Train	Test	
		Same	Other
Tomato (10 classes)	38,648	4,295	13,956
Strawberry (4 classes)	16,859	1,874	2,307
Cucumber (10 classes)	82,845	9,206	14,760
Eggplant (6 classes)	30,645	3,405	3,042
Total	168,997	18,780	34,065

Table 1: Number of dataset for four crop diseases.

2 Dataset

In this study, we used a total of 221,842 images of 20 diseases and healthy leaves from four crops (tomato, strawberry, cucumber, and eggplant). The summary of the data used in the study is shown in Table 1 (more details can be found in Table 3 in the Appendix). The diseases we dealt with are 1: Powdery Mildew (PM), 2: Gray Mold (GM), 3: Anthracnose (AN), 4: Cercospora Leaf Mold (CLM), 5: Leaf Mold (LM), 6: Late Blight (LB), 7: Downy Mildew (DM), 8: Corynespora Leaf Spot (CLS), 9: Corynespora Target Spot (CTS), 10: Leaf Spot (LS), 11: Fusarium Wilt (FW), 12: Gummy Stem Blight (GSB), 13: Verticillium Wilt (VW), 14: Bacterial Wilt (BW), 15: Bacterial Spot (BS), 16: Bacterial Canker (BC), 17: Cucurbit Chlorotic Yellows Virus (CCYV), 18: Mosaic Diseases (MD), 19: Melon Yellow Spot Virus (MYSV), 20: Yellow Leaf Curl (YLC), and 0: Healthy (HE). Images used in this experiment were prepared by experts who inoculated each plant with the disease or infected it via insects to obtain accurate training labels.

To avoid co-infection, we grew our crops in isolated facilities under strict control. Therefore, the dataset includes not only typical symptoms but also images of early symptoms of infection. These images were taken under the condition of the leaves being roughly in the center of the picture.

3 Experiments

We conducted two experiments based on a large number of leaf images of four different crops collected by agricultural institutions from 24 prefectures in Japan.

- Experiment 1: Assessing the need for separation of training and test data
- Experiment 2: Assessing the effect of pre-detection of ROI on diagnostic performance

For our experiments, we used EfficientNet-B4 (Tan and Le 2019), which is a sophisticated CNN model and has been reported to have excellent performance on image classification in recent years. The input image size was set to 512×512, and basic random online data augmentations were performed. They are horizontal flip, vertical flip, random rotation with a step size of 20, cropping to include more than 80% of the image, and brightness change. As for the evaluation criteria, we used micro accuracy and F1-score.

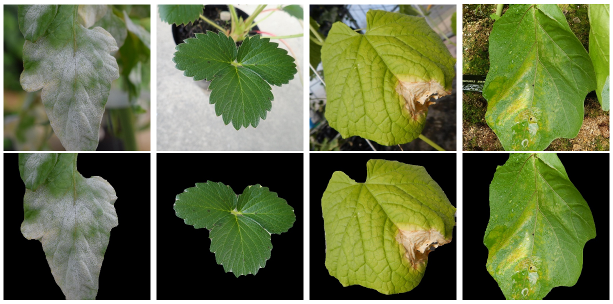


Figure 2: Examples of original (top row) and leaf images with the background removed using AOP (bottom row)

3.1 Effect of separation of training and evaluation data

First, we investigated the extent to which the potential similarity of the images would result in overfitting of the model as Experiment 1. In order to compare the diagnostic performance of images taken in different locations, we prepared two test sets: one with images taken in the same field as the training data, and one from different fields. The farms were selected so that the number of images was as large as possible while meeting the requirements. As shown in Table 1, a total of 168,814 images were used for training, with 18,902 test images taken in the same field as the training image, and exclusive 34,065 test images taken in the other fields.

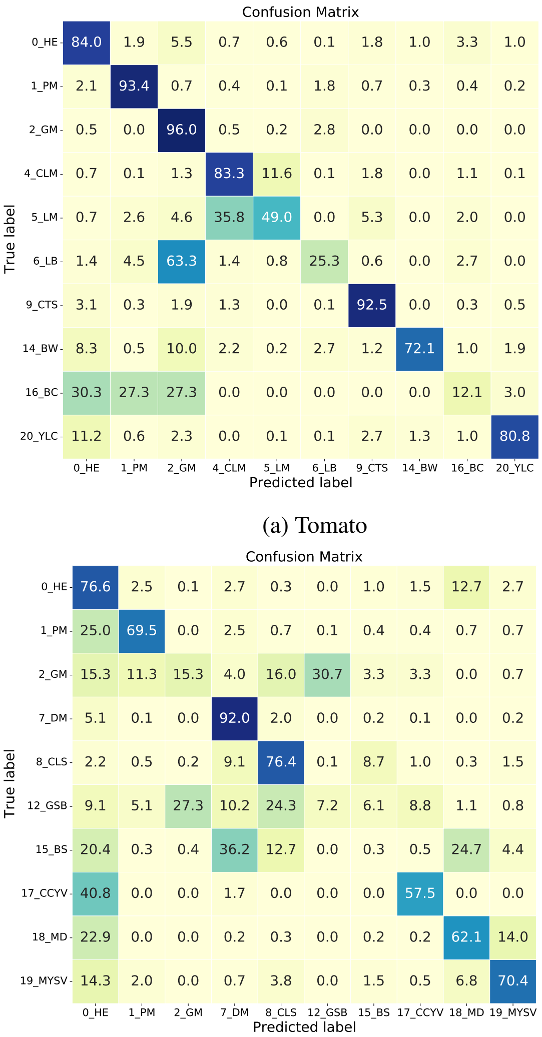
3.2 Effect of the ROI detection

Second, we verified the effectiveness of the pre-detection of the region of the interest as Experiment 2. We used AOP (Saikawa et al. 2019), a high-quality leaf region detection method based on pix2pix (Isola et al. 2017). Using a set of original images and the corresponding mask images, AOP is trained to segment the leaf region from input images. Figure 2 shows examples of ROI detection, i.e., backgrond removal for four crops. We confirmed that appropriate segmentation was achieved for most cases by visual evaluation. For more details of AOP, see Saikawa et al. (2019). To train the AOP model, a total of 28,661 leaf from the training images categories in Table 1 were used: 9,725 tomatoes, 2,388 strawberries, 7,881 cucumbers, and 8,667 eggplants images. The trained AOP model then was used to segment out the leaf regions as the ROI in the test data prior to diagnosis.

4 Results

4.1 Need for separation of evaluation data

Table 2 shows the comparison of the disease diagnostic performance when the test image is the same as and different from the training image. When images of the same field were evaluated, both accuracy and F1-score were about 99%, which is similar to what has been reported in other papers. On the other hand, however, we observed that the covariate shift between training and test images significantly degraded the diagnostic performance for many disease classes in all crops. Figure 3 shows the confusion matrices for (a) tomato and (b) cucumber, where the test images were collected at different sites than the training images. Those for strawberry and eggplant are in Figure 4 in the Appendix.



(a) Tomato

(b) Cucumber

Figure 3: Confusion matrices of (a) tomato and (b) cucumber, where the test images were collected at different sites than the training images.

4.2 The effect of ROI detection

The evaluation of pre-detection using AOP is also shown in Table 2. Here, (+ ROI) in the bottom line indicates the diagnosis with ROI. From the diagnostic results, there is no significant improvement in accuracy and F1-score, which shows that background information has little effect on the final performance in this experiment.

5 Discussion

5.1 The impact of test data independence

From Experiment 1, it was confirmed that the diagnostic performance of all four crops decreased when the evalua-

Metric	Test Field	Tomato (10 class)	Strawberry (4 class)	Cucumber(10 class)	Eggplant (6 class)
Micro Acc. [%]	same farm	99.1	99.6	98.8	98.9
	other farm	84.1	88.3	66.4	81.6
	other farm (+ROI)	83.6	86.5	66.0	82.2
Macro F1. [%]	same farm	99.0	99.5	98.2	98.9
	other farm	65.2	87.6	49.6	76.3
	other farm (+ROI)	66.4	85.0	51.2	76.0

Table 2: Comparison of diagnostic performance. The (+ROI) indicates diagnosing with background removal.

tion data was collected from different farms than the training data. Since the appearance of images varies greatly from farm to farm due to differences in symptom presentation, geometry, and a variety of other conditions, covariate shifts can be significant, resulting in poor diagnostic performance.

One obvious reason is that regular CNNs are not able to deal with scaling in nature, so they are strongly affected by differences in geometry if they have not acquired enough diversity in their training data. Among all crops, the performance gap was pronounced for tomato and cucumber, which have many disease species. This is because diseases with similar symptoms to each other are inherently more difficult to identify. This gap also tends to be larger when the number of farms from which training data is provided is small. Increasing the amount of training data at each site can greatly contribute to improving handling of diversity within that field, but can only have a limited effect on improving generalizability to very large domain shifts across fields. Therefore, for intrinsic performance improvement, it is necessary to collect training data in as many sites as possible, rather than collecting a lot of training data in a limited number of sites.

On the other hand, when the evaluation set was from the same population as the training set, high diagnostic identification performance was achieved, as shown in existing studies. In this case, it is not a practical but an *inauthentic* performance that can only be achieved in cases similar to the training data. Evaluations using more data and newer classification models than previously reported studies have again confirmed this tendency. Therefore, it is natural to assume that the results of the previous papers are due to overfitting caused by training on the same field data, which is yielding high performance.

5.2 The effect of ROI detection

In this experiment, the pre-detection of ROI had little effect on the improvement of diagnostic performance, which was different from Saikawa et al. (2019). Although they used nearly 36,000 training images of cucumbers, their VGG-16 model (Simonyan and Zisserman 2015) trained on 224×224 low-resolution images may not have detected enough disease features. Therefore, we believe that their CNN model overfitted to the background image, and some accuracy improvement was obtained by background removal. On the other hand, in this experiment, we trained our EfficientNet model with larger training data (about 83,000 images for cucumbers) and higher resolution images (512×512). This may have made it easier for the classifier to detect symptoms even in the presence of a background. However, as the results show, the effects of domain shift are still present.

In general, it is very important to focus on cue regions, or ROIs, in identification and classification problems, as evidenced by the recent results of attention mechanisms in machine learning. The results show that in situations where a certain resolution is obtained, the background is not the main cause of this overfitting, but factors present in the ROI (such as shooting distance, light conditions, camera quality, development stage, etc.) affect the diagnostic performance. ROI detection did not contribute to the performance improvement in this experiment. This can be interpreted not as being because the detection was meaningless, but rather there were too many differences between the domains. Although improving the intrinsic performance for plant diseases is not the main purpose of this paper, appropriate pre-processing to reduce the covariate shift, such as standardizing the distance between the camera and the diagnostic target, normalizing the color, and applying domain adaptation techniques, can improve the accuracy. We believe that ROI detection can be an effective method if we can eliminate too large a covariate shift, and we will investigate this in the future.

6 Conclusion

We used a large and reliable image dataset taken from four crop fields to investigate two issues: (1) the importance of source independence for training and evaluation data, and (2) the importance of ROI pre-detection. As a result, we have shown how the evaluations in many previous studies are inadequate because of the potential similarities in images from the same field, and because of the impact of inappropriately high apparent ratings when these data are taken from the same source. The characteristics of the images in each field may vary greatly, and this affects not only the background of the image but also the ROI regions with symptoms. Therefore, ROI detection alone cannot solve the problem of performance degradation due to field differences, and measures to reduce such covariate shift are necessary.

Acknowledgments

This work was supported by Japanese Ministry of Agriculture, Forestry and Fisheries (MAFF) commissioned project study on “Development of pest diagnosis technology using AI” (JP17935051).

References

Boulent, J.; Foucher, S.; Théau, J.; and St-Charles, P.-L. 2019. Convolutional neural networks for the automatic identification of plant diseases. *Frontiers in plant science*, 10: 941.

Brahimi, M.; Arsenovic, M.; Laraba, S.; Sladojevic, S.; Boukhalfa, K.; and Moussaoui, A. 2018. Deep learning for plant diseases: detection and saliency map visualisation. In *Human and machine learning*, 93–117. Springer.

Brahimi, M.; Boukhalfa, K.; and Moussaoui, A. 2017. Deep learning for tomato diseases: classification and symptoms visualization. *Applied Artificial Intelligence*, 31(4): 299–315.

Cap, Q. H.; Tani, H.; Kagiwada, S.; Uga, H.; and Iyatomi, H. 2021. LASSR: Effective super-resolution method for plant disease diagnosis. *Computers and Electronics in Agriculture*, 187: 106271.

Cap, Q. H.; Uga, H.; Kagiwada, S.; and Iyatomi, H. 2020. Leafgan: An effective data augmentation method for practical plant disease diagnosis. *IEEE Transactions on Automation Science and Engineering*.

Durmuş, H.; Güneş, E. O.; and Kırcı, M. 2017. Disease detection on the leaves of the tomato plants by using deep learning. In *2017 6th International Conference on Agro-Geoinformatics*, 1–5. IEEE.

FAO. 2020. Protecting Plants, Protecting Life. *Food and Agriculture Organization of the United Nations*.

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

Fuentes, A.; Yoon, S.; Kim, S. C.; and Park, D. S. 2017. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, 17(9): 2022.

Fujita, E.; Kawasaki, Y.; Uga, H.; Kagiwada, S.; and Iyatomi, H. 2016. Basic Investigation on a Robust and Practical Plant Diagnostic System. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 989–992.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kanno, S.; Nagasawa, S.; Cap, Q. H.; Shibuya, S.; Uga, H.; Kagiwada, S.; and Iyatomi, H. 2021. PPIG: Productive and Pathogenic Image Generation for Plant Disease Diagnosis. In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 554–559.

Liu, J.; and Wang, X. 2020. Early recognition of tomato gray leaf spot disease based on MobileNetv2-YOLOv3 model. *Plant Methods*, 16: 1–16.

Ma, J.; Du, K.; Zheng, F.; Zhang, L.; Gong, Z.; and Sun, Z. 2018. A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network. *Computers and electronics in agriculture*, 154: 18–24.

Mithu, M. A.; Momo, S. I.; Hasan, M.; Rahman, K. M.; Sattar, A.; et al. 2022. Pumpkin Leaf Disease Detection: Convenience of CNN Over Traditional Machine Learning in Terms of Image Classification. In *Smart Systems: Innovations in Computing*, 347–357. Springer.

Mohanty, S. P.; Hughes, D. P.; and Salathé, M. 2016. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7: 1419.

Saikawa, T.; Cap, Q. H.; Kagiwada, S.; Uga, H.; and Iyatomi, H. 2019. AOP: an anti-overfitting pretreatment for practical image-based plant diagnosis. In *2019 IEEE International Conference on Big Data (Big Data)*, 5177–5182. IEEE.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Suwa, K.; Cap, Q. H.; Kotani, R.; Uga, H.; Kagiwada, S.; and Iyatomi, H. 2019. A comparable study: Intrinsic difficulties of practical plant diagnosis from wide-angle images. In *2019 IEEE International Conference on Big Data (Big Data)*, 5195–5201.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114. PMLR.

Tani, H.; Kotani, R.; Kagiwada, S.; Uga, H.; and Iyatomi, H. 2018. Diagnosis of Multiple Cucumber Infections with Convolutional Neural Networks. In *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–4.

Toda, Y.; and Okura, F. 2019. How convolutional neural networks diagnose plant disease. *Plant Phenomics*, 2019.

Wang, G.; Sun, Y.; and Wang, J. 2017. Automatic image-based plant disease severity estimation using deep learning. *Computational intelligence and neuroscience*, 2017.

Yu, H.-J.; and Son, C.-H. 2020. Leaf Spot Attention Network for Apple Leaf Disease Identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 229–237.

Zekiwos, M.; Bruck, A.; et al. 2021. Deep Learning-Based Image Processing for Cotton Leaf Disease and Pest Diagnosis. *Journal of Electrical and Computer Engineering*, 2021.

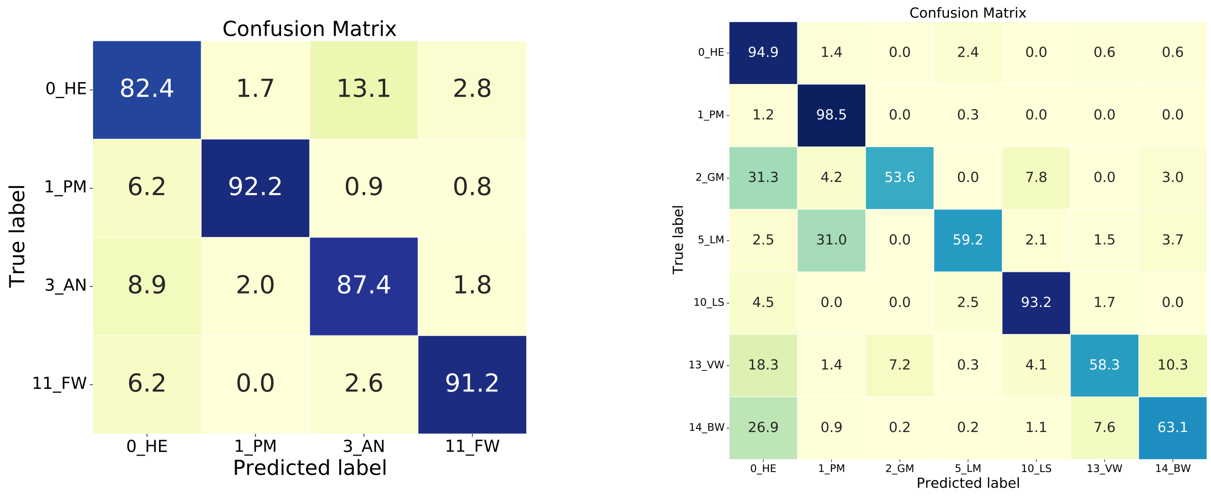
Zhang, J.; Rao, Y.; Man, C.; Jiang, Z.; and Li, S. 2021. Identification of cucumber leaf diseases using deep learning and small sample size for agricultural Internet of Things. *International Journal of Distributed Sensor Networks*, 17(4): 15501477211007407.

Appendix

Table 3 shows the details of the datasets that could not be included due to space limitations. Here, the number in parentheses is the number of plots for which images were collected. Figure 4 is the confusion matrices for (a) strawberry and (b) eggplant.

ID / Name	Tomato			Strawberry			Cucumber			Eggplant		
	Train	Test		Train	Test		Train	Test		Train	Test	
		Same	Other		Same	Other		Same	Other		Same	Other
0_HE	7,307 (6)	813	2,994	9,440 (6)	1,032	578	14,383 (5)	1,633	5,309	9,784 (4)	1,122	831
1_PM	4,009 (5)	481	4,250	1,770 (6)	182	893	6,211 (4)	667	1,898	6,782 (4)	772	861
2_GM	8,427 (3)	900	571				581 (1)	62	150	920 (1)	104	166
3_AN				3,333 (5)	368	609						
4_CLM	3,504 (3)	370	1,809									
5_LM	2,462 (3)	261	151							2,363 (2)	260	326
6_LB	1,841 (1)	198	640									
7_DM							6,244 (3)	709	1,260			
8_CLS							6,864 (4)	701	1,813			
9_CTS	1,547 (2)	185	1,350									
10_LS										4,847 (3)	505	118
11_FW				2,316 (5)	292	227						
12_GSB							1,320 (2)	163	374			
13_VW										2,856 (2)	320	290
14_BW	1,972 (4)	214	412							3,093 (3)	322	450
15_BS							3,914 (2)	414	1,147			
16_BC	3,535 (1)	404	33									
17_CCYV							5,339 (1)	630	179			
18_MD							24,209 (1)	2,651	1,626			
19_MYSV							13,780 (4)	1,576	1,004			
20_YLC	4,044 (4)	469	1,746									
Total	38,648	4,295	13,956	16,859	1,874	2,307	82,845	9,206	14,760	30,645	3,405	3,042

Table 3: Details of the dataset numbers for four crop diseases. (*) indicates the number of fields. For example, in 0_HE of tomato, the number of fields is 6.



(a) Strawberry

(b) Eggplant

Figure 4: Confusion matrices of (a) strawberry and (b) eggplant, where the test images were collected at different sites than the training images.