Annotation Consistency Analysis for Plant Disease Detection

Jiuqing Dong
Department of Electronic Engineering,
Jeonbuk National University
South Korea
jiuqingdong1996@jbnu.ac.kr

Alvaro Fuentes
Core Research Institute of Intelligent
Robots, Jeonbuk National University
South Korea
afuentes@jbnu.ac.kr

Sook Yoon
Department of Computer Engineering,
Mokpo National University
South Korea
syoon@mokpo.ac.kr

Taehyun Kim
National Institute of Agricultural Sciences,
Wanju
South Korea
thkim8205@korea.kr

Dong Sun Park
Department of Electronic Engineering,
Jeonbuk National University
South Korea
dspark@jbnu.ac.kr

ABSTRACT

Object detection models have become the current tool of choice for plant disease detection in precision agriculture. In specific domains, the definition of objects is completely different from typical datasets, so acquiring precise annotations is no easy task in professional and natural contexts. Annotation inconsistency is one of the main factors plaguing the performance of object detection models. In this work, we define five different types of inconsistencies in the annotation process and investigate the severity of the impact of inconsistent labels on model's performance. We also conduct an interpretability study of the inconsistency analysis by using class activation maps. Overall, this data-centric quantitative analysis helps us to understand the significance of annotation consistency, which provides practitioners with experience in annotation on plant disease detection. Our work encourages researchers to pay more attention to annotation consistency.

KEYWORDS

Plant disease detection, Inconsistency bounding box, Noise analysis.

1 INTRODUCTION

Recently, image acquisition is becoming easier and easier in the agricultural field because of the increased use of cameras and sensors. Intelligent applications based on agricultural images have been widely used in many aspects of agriculture, especially in plant disease and pest detection [1-4]. Deep learning technology has been playing a dominant role in various detection tasks. This technology can potentially reduce the negative impacts of plant diseases by promptly estimating the damage using non-intrusive sensors such as RGB cameras. Deep learning techniques rely on well-annotated datasets. Nevertheless, acquiring an accurately annotated dataset is not always feasible due to several factors. For example, practitioners without computer vision knowledge lack experience on how to annotate high-quality boxes, while annotators without domain knowledge are also difficult to annotate accurate object boxes. Due to these practical challenges, the actual labels often deviate from the ideal value, which leads to labels that are inconsistent with the instances.

Even though extensive works of deep learning techniques under class noise exist [5-8], it mainly focuses on computer vision datasets such as MS-COCO [9], PASCAL VOC [10], and ImageNet [11] rather than domain-specific datasets. In some domains, the definition of an object is significantly different from generic objects in these typical datasets, thus bringing annotation ambiguities. The labels provided by domain experts are always with high quality. However, it is time-consuming and expensive to involve domain experts to do the annotation. Moreover, it is non-scalable if a considerable number of labels are needed. Many applications have started to use crowd-sourcing technology to achieve labels. For example, some websites ask an operator if an image contains a specific object to determine that the operator is not an autonomous robot. That means many netizens are contributing annotations even if they are not computer vision experts. A common situation is that after experts provide a good annotation scheme, the dataset is assigned to annotators for annotation. However, the quality of those non-expert annotations is hard to control. We believe that people may give incorrect labels when they repeat the same work for a long time because they have already lost patience. This non-expert label noise is a common issue, especially in pervasive and ubiquitous computing or lifelong learning. Up to now, the existing literature lacks a specific study on inconsistent labels in plant disease detection.

Consequently, we aim to study the impact of inconsistent annotations on model performance, thereby providing insights into the follow-up practitioners in plant disease detection. Our main contributions are summarized as follows:

1. We define five types of inconsistencies to describe annotation inconsistencies and perturb clean bounding boxes to simulate a noisy label set. Then, we investigate the severity of the impact on

the model's performance. It is the first quantitative study of annotation inconsistency in plant disease detection.

2. We explain the impact of inconsistent labels on the feature extractor through visualization techniques, which can help us analyze the logic behind recognizing data with noise.

2 Related Work

2.1 Annotation consistency analysis

In different works, inconsistency might be referred to different definitions. In some works [12, 13], it is referred to outliers and anomaly. In Research [14] and noise filtering in a medical domain [15], label noise means those instances which disproportionately increase the model complexity. Frénay [16] gives a comprehensive survey on different types of label noise. In their work, label noise is considered to be the observed labels which are classified incorrectly.

Zhu X et al. [17] differentiate noise into two categories: class noise and attribute noise. Then they analyze their impacts on the model's performance separately. Flatow D et al. [18] examine how sensitive the convolutional neural network (CNN) model is to noise in the training set, particularly when the training set contains mislabeled or subjectively-labeled examples. They took an approach to simulate inconsistencies in the training set. Nazari Z et al. [19] evaluated the class noise impact on the performance of three widely used machine learning algorithms. Xu M et al. [20] studied the missing instance-level label problem in object detection. Li Y et al. [21] analyzed crop pest and disease datasets regarding data quality. They found that high-quality data can bring about 10% to 20% performance improvement. Algan G et al. [22] make a comprehensive survey of methodologies centered explicitly around deep learning in the presence of noisy labels. Nevertheless, most works focus on classification and object detection tasks on typical datasets rather than domain-specific ones.

In plant disease detection, the definition of the object is significantly different from generic objects in MS-COCO and PASCAL VOC, because the symptoms appear in a specific part of the leaf. In contrast, almost all bounding boxes in the MS-COCO dataset cover the complete target except for the incompletely displayed objects. However, annotation consistency will suffer from the degenerated label set, including wrongly assigned class labels and inaccurate, redundant, and missing bounding boxes, which can impact interpretations of the data, models created from the data, and decisions made based on the data. Although CNNs are relatively robust to class noise, it is worthwhile to investigate in a more comprehensive study of inconsistency.

2.2 Deep-learning Technics

CNN-based architectures in deep learning are proven to be the best in learning representations and solving complex computer vision and general artificial intelligence problems. Researchers revisited the architecture of CNNs throughout the recent years and proposed a series of famous works. A typical view is that the feature extractor's capacity benefits from several factors, such as a wider and deeper network, a higher resolution input, and more powerful hardware [23, 24]. An interesting observation in [23]: with a similar architecture, by scaling the network depth, width, and input resolution, the image classification performance on ImageNet only improved by 2.4%, while the number of parameters increased from 22M to 208M. It is unwise to spend a considerable cost for a slight improvement in practical applications.

Data-centric deep learning approaches attempt to improve performance by analyzing the relationship between data quality, quantity, and networks. Also, the community devotes itself to detecting and fixing noise generated by the data collection and annotation. Up to now, it remains an open issue. For example, a poor instance initialization could render failure during training [25]. Data annotation consistency is crucial in object detection, affecting feature extraction performance from objects. Therefore, in the field of plant disease detection, it is necessary to study the relationship between data quality and model performance.

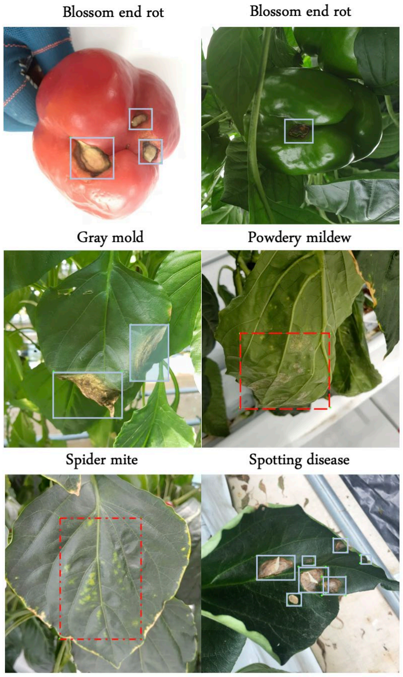

Figure 1: Samples of each kind of disease and annotations

3 Materials and Methods

3.1 Dataset

In this paper, paprika is used as a target plant for annotation analysis in plant disease detection, and its target diseases are four leaf diseases, gray mold, powdery mildew, spider mite, and spotting disease, and one fruit disease, blossom end rot. The dataset consists of 5928 paprika images. The samples of leaf diseases were collected in the greenhouse, while fruit disease was collected in the greenhouse and laboratory. Our dataset is much more complex because more than one disease may appear on a single leaf. In addition, we believe that the annotation strategies employed by different symptoms also should be different. Specifically, as for blossom end rot, gray mold, and spotting disease, there is a distinguishable boundary between the abnormal and normal areas. We can implement refined annotation for these three diseases. On the contrary, as for powdery mildew and spider mite, it is hard to demarcate the boundaries of instances of those classes to annotate precisely. Therefore, we use a larger bounding box to annotate the two diseases. Samples and annotations are shown in Fig. 1. More details are shown in Table 1.

Table 1: Details of datasets. We use A+B to denote the total number of labels, where A contains clear instances and B contains blurred instances (redundant labels).

Category	Images	Instances
Blossom end rot	1183	1563+39
Gray mold	441	670+47
Powdery mildew	416	421+25
Spider mite	420	1447+93
Spotting disease	3468	9741+347
Total	5928	13842+551

3.2 Annotation Inconsistency

Fully supervised object detection requires that each instance should be annotated by an accurate bounding box. Intuitively, we expect the selected instance covers the actual object as tight as possible. However, affected by empirical and systematic errors, the quantity and attributes of the factual bounding boxes in the label set cannot accurately represent all instances. In this paper, we define five different types of inconsistency existing in real scenarios according to the practical task of plant disease detection into quantity-based inconsistency and attribute-based inconsistency. To our best knowledge, it is the first time to make an analysis on inconsistent labels for plant disease detection. Fig. 2 shows examples of several different types of inconsistency.

3.2.1 Quantity-Based Inconsistency. Due to subjective or objective factors, the number of bounding boxes deviates from the correct during the annotation process. We call quantity-based inconsistency.

Redundant Label. A redundant label is used to describe an instance that should not have been labeled or is labeled multiple times. In a plant disease detection task, some blurred disease areas are generated due to the focus problem of the camera. These blurry disease areas maybe are annotated. In this work, we refer to a bounding box that includes a blurry instance as a redundant label. An example of a redundant label is shown in Fig. 2(A).

Missing Label. Fully supervised object detection assumes that every example belonging to the target class should be annotated. A missing label is used to describe an instance that is ignored. The missing label profoundly affects the performance of Fully supervised object detection methods. Thus, we also focus on studying the relationship between missing labels and the model's performance on plant disease detection. An example of a missing label is shown in Fig. 2(B).

3.2.2 Attribute-Based Inconsistency. Bounding boxes have several representations in object detection tasks. For example, in YOLO-based methods, (c, x, y, w, h) denotes the attribute of the bounding box, where (x, y) denotes the coordinates of the center point, and (w, h) represents the width and height of the bounding box. According to these three aspects, we define three kinds of attribute inconsistencies: class noise, position inconsistency, and size inconsistency.

Class noise. Due to random factors and human error, a common noise is assigning an instance with a wrong category, which we call class noise. In this work, we only discuss the impact of random class noise on the model's performance. An example of class noise is shown in Fig. 2(C).

Position inconsistency. The bounding box's centric coordinates are crucial for the model to regress the target's position. In this work, position inconsistency refers to the degree of centric coordinate deviation between noisy ground truth and original ground truth. An example of position noise is shown in Fig. 2(D).

Size inconsistency. Each disease area needs to be allocated a reasonably sized bounding box which determines how well the box fits the object. In this paper, size inconsistency is used to describe that a bounding box does not match the size of the instance. An example of size noise is shown in Fig. 2(E).

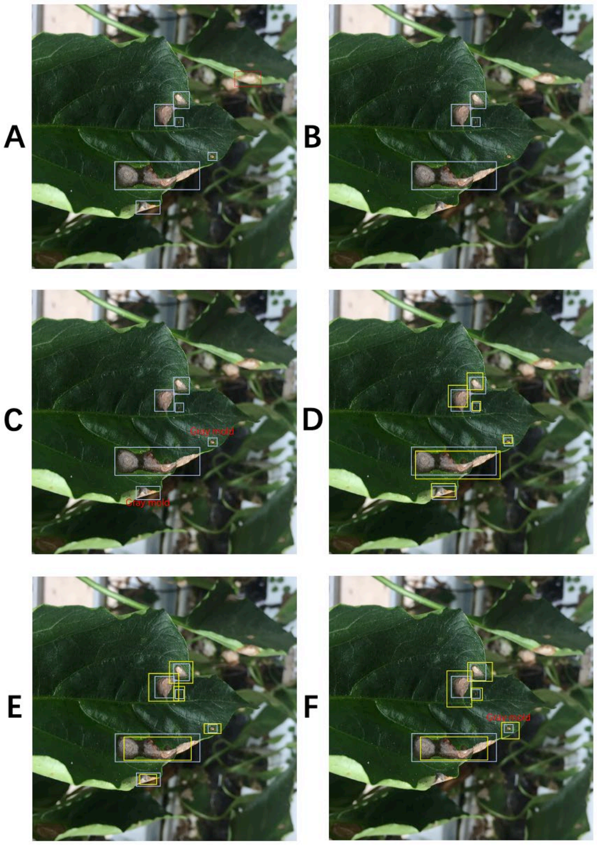

Figure 2: Different types of annotation inconsistency. Blue bounding boxes, yellow boxes, and red dashed boxes represent the ground truth, inconsistent labels, and a blurry instance, respectively. (A). Redundant labels. (B). Missing labels. (C). Class noise. (D). Position inconsistency. (E). Size inconsistency. (F). Mixed attribute inconsistency.

3.3 Evaluation Metrics

We evaluate the performance of the bounding box detector using the following metrics:

Mean Average Precision score (mAP): mAP is the area under the precision-recall curve calculated for all classes.

$$AP = \frac{1}{11} \sum_{r\in[0,0.1,\dots,0.9,1]} P(r) \qquad (1)$$

$$P(r) = \max_{\tilde{r}:\tilde{r}\geq r} p(\tilde{r}) \qquad (2)$$

where, $P(r)$ is the maximum precision for any recall values greater than r, and $p(\tilde{r})$ is the measured precision at recall $\tilde{r}$.

4 Experiments and Results

4.1 Implements Settings

We implement experiments on paprika disease dataset, which includes five annotated disease classes. Our dataset was divided into 80% training set, 10% validation set, and 10% testing set in all experiments. Training is proceeded on the training set, and the evaluation is performed on the validation set. When the experiments reach the expected results, the final evaluation is done on the testing set. YOLO-v5 is a representative work of real-time detection in the industry. Therefore, we evaluate the effectiveness of the dataset with different inconsistent levels on YOLO-v5 model. The training parameters are consistent with the original YOLO-v5-extra-large model, which was trained and tested on 3 GTX 3090 GPUs and implemented in PyTorch 1.10.1.

4.2 Annotation Inconsistency

In this part, we analyze the impact of inconsistency on the model's performance, including redundant labels, missing labels, class noise, position inconsistency, and size inconsistency. We treat the annotated boxes of blurred instances as redundant labels. The number of blurred instances for each disease is shown in Table 1. Table 2 shows the comparison of model's performance on the label set with and without the redundant label. As for missing labels and class noise, we randomly perturb or remove the bounding box, where the probability ranges from 5% to 20%. In YOLO-based methods, (c, x, y, w, h) denote attribute of a bounding box, where (x, y) denotes the coordinates of the center point, and (w, h) represents the relative width and height of the bounding box. We use I to indicate the deviation degree of position inconsistency and size inconsistency, where I range from 5% to 20%. The perturbing method is shown in Eq. 3. In this way, we can analyze whether the different types of inconsistency have the same impact on each disease. Finally, we add three attribute noises simultaneously, which we call mixed attribute inconsistency.

$$\begin{cases} x_{inc} = x \pm 0.5 * \Delta x * w, & y_{inc} = y \pm 0.5 * \Delta y * h \\ w_{inc} = w \pm \Delta w * w, & h_{inc} = h \pm \Delta h * h \end{cases} \qquad (3)$$

where $x_{inc}, y_{inc}, w_{inc}, h_{inc}$ denote the attribute of a bounding box after perturbing. $\Delta x, \Delta y, \Delta w, \Delta h$ are perturbation coefficients. While I equal to 20%, the $\Delta x, \Delta y, \Delta w, \Delta h$ are in the range of [-0.2,0.2]. Note that Eq.3 is performed on every bounding box in the training data.

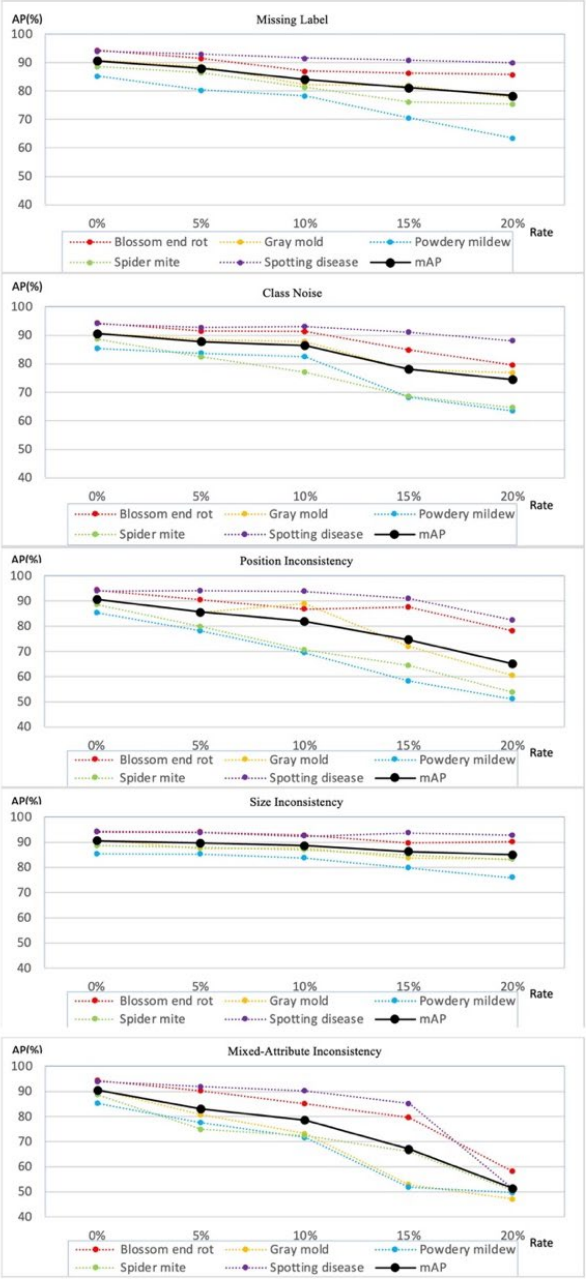

Figure 3: Sensitivity of YOLO-v5 extra-large model to different types of inconsistency.

As shown in Fig. 3, the training accuracy linearly decreases with the increasing level of annotation inconsistency but varies among different types of inconsistency. Our observations are consistent with those observed in related work on noise. In the five types of inconsistency, positional noise devastated the model's performance. Conversely, the model is less sensitive to size inconsistency and missing labels, but the accumulated size inconsistency and missing label can still cause severe problems for the model. From the decreasing tendency, diseases with fewer instances are more susceptible to inconsistency. For example, when the mixed attribute inconsistency continued to increase, the average precision of gray mold, powdery mildew, and spider mite diseases dropped from around 90% to half.

Table 2: Comparison of with or without redundant label.

Category	without Redundant Noise	with Redundant Noise
Blossom end rot	94.3 %	91.2 %
Gray mold	90.8 %	84.8 %
Powdery mildew	85.3 %	80.9 %
Spider mite	88.6 %	82.1 %
Spotting disease	94.0 %	91.5 %
mAP	90.6 %	86.1 %

4.3 Visualization

Eigen-CAM [26] takes advantage of the principal components to improve the weights, which is proven to be a very efficient and convenient visualization method. It can work with all CNNs without modifying layers or retraining models. More conveniently, Eigen-CAM can visualize the activation map of any layer in the neural network, which helps us to understand what the CNN learns. Fig. 4 shows the visualization results of YOLO-v5-x model learning on datasets with different noise levels. To further explore the relationship between prediction results and activation maps, we remove heatmap data outside the bounding boxes, and scale the heatmaps inside every bounding box. Then we can obtain the final CAM with category information (Refer to the last column of Fig. 4).

We observed that the dataset with noisy labels is less accurate for localizing disease regions than the clean dataset. In Case 1, affected by inconsistent labels, the suspective area is difficult to activate, even if the disease has been detected. In Case 2, the focus on the suspective area diminishes as the inconsistency level increases. Final results did not appear to be affected, possibly due to a larger disease area and the symptoms were completely distinguishable from the leaves. Moreover, In Case 3, with the increasing inconsistent level, more negative false bounding boxes appeared (Refer to the last column in Fig. 4).

4.4 Discussion

Annotation consistency is an essential indicator in evaluating the quality of data. Li Y et al. [21] concluded that limited good data could beat a lot of bad data. In our work, we found that datasets with inconsistent labels (bad data) caused a much more significant drop in performance than missing labels (limited good data), which is consistent with their conclusion. Fig. 3 shows that the model's performance decreased to around 50%, with the mixed attribute inconsistency level increasing to 20%. In contrast, with 20% of the labels missing, the model's performance is still above 75%. Therefore, we argue that an inconsistent label may significantly impact the network model more than a missing label. Besides, the results in [25] show that the CNN method can achieve the same performance as a clean dataset on a dataset with 10% label noise with advanced annotation correction techniques. Nonetheless, we still recommend that annotators strive to improve the consistency and accuracy during the annotation process rather than using correction techniques directly.

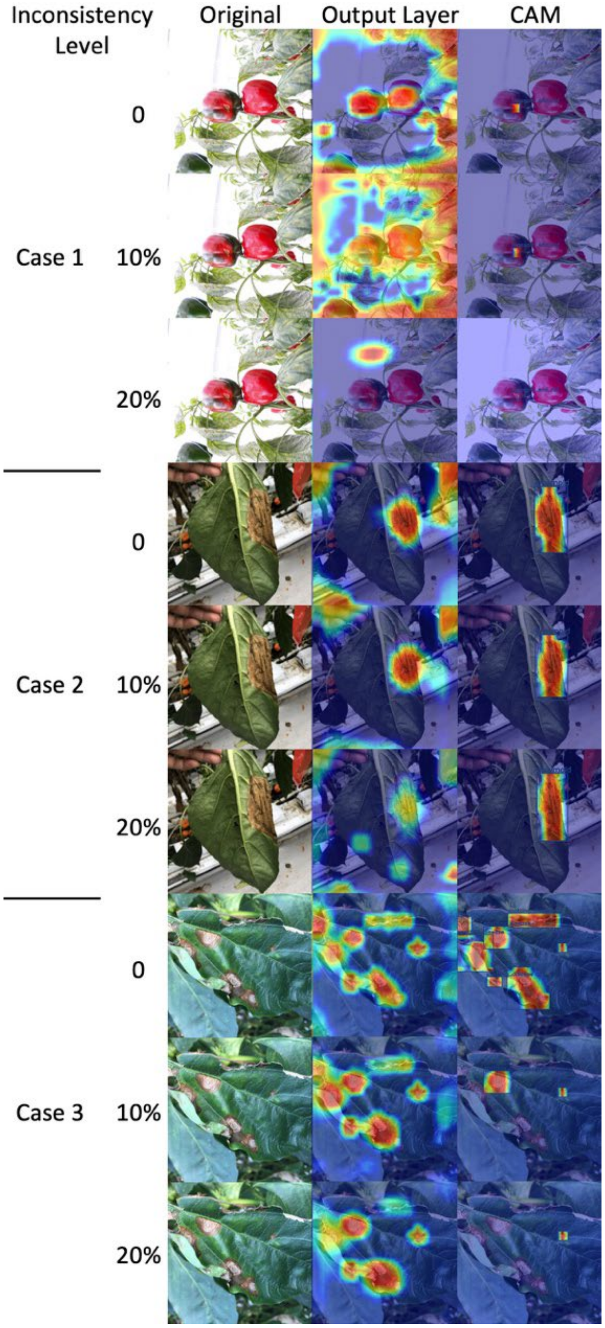

Figure 4: Class activation maps for different levels of inconsistency. We visualize the output layer of the YOLO-5-extra-large model and CAM with bounding boxes.

Fig. 5 shows the sensitivity of each disease to noisy data. Inconsistent labels had minimal impact on spotting disease. It may be due to a large number of instances, which greatly improves the anti-interference ability of the network. Another possible reason is that the symptoms of spotting disease are easy to distinguish, which enhanced identification capabilities of CNNs. Conversely, powdery mildew and spider mite suffered impacts from inconsistent labels more, because their symptoms are hard to localize.

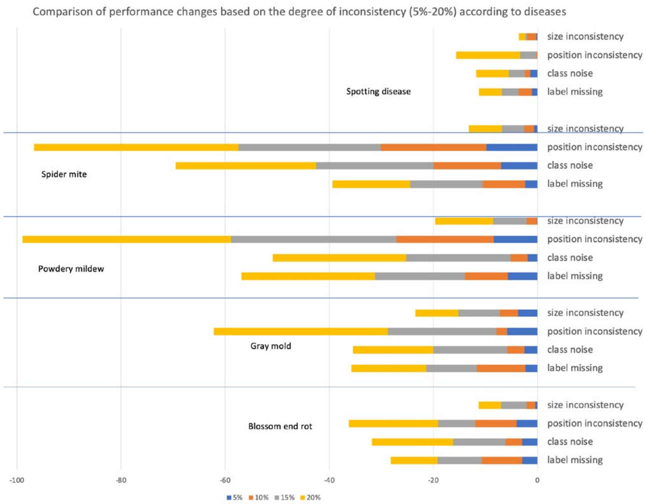

Figure 5: Comparison of performance changes based on the degree of inconsistency (5%-20%) according to diseases.

5 CONCLUSIONS

The detection of plant diseases using digital images is a challenging task. Analyzing annotation inconsistency in advance is necessary for plant disease detection. Experiments demonstrated that the impacts of inconsistency are severe in many circumstances. Compared with other kinds of inconsistency, position inconsistency is more damaging than class noise. The size inconsistency is usually less harmful but still could lead to a slight reduction in the performance of learning algorithms. In addition to the above work, we also emphasized the interpretability of our methods. With these conclusions, instead of adopting some 'blind' noise handling mechanisms, interested readers can design their own inconsistency handling approaches to enhance data quality from their perspectives. Due to the enormous annotation cost, our work provides guidelines to some extent but still has limitations. There is currently a lack of research on inconsistency handling mechanisms in plant disease detection. Our future work will focus on automatically repairing the errors in the data labels and turning the low-quality data into high-quality data.

ACKNOWLEDGMENTS

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 Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) and Korea Smart Farm R&D Foundation(KosFarm) through Smart Farm Innovation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA) and Ministry of Science and ICT(MSIT), Rural Development Administration(RDA)(421005-04).

REFERENCES

[1] Fuentes, A., et al., A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors, 2017. 17(9): p. 2022.

[2] Fuentes, A., et al., Open Set Self and Across Domain Adaptation for Tomato Disease Recognition With Deep Learning Techniques. Frontiers in Plant Science, 2021. 12.

[3] Fuentes, A.F., et al., High-performance deep neural network-based tomato plant diseases and pests diagnosis system with refinement filter bank. Frontiers in plant science, 2018. 9: p. 1162.

[4] Dong J, Fuentes A, Yoon S, et al. Towards Improved Performance on Plant Disease Recognition with Symptoms Specific Annotation[J]. Smart Media Journal, 2022, 11(4): 38-45.

[5] Bernhard, M. and M. Schubert, Correcting Imprecise Object Locations for Training Object Detectors in Remote Sensing Applications. Remote Sensing, 2021. 13(24): p. 4962.

[6] Li, J., et al., Towards noise-resistant object detection with noisy annotations. arXiv preprint arXiv:2003.01285, 2020.

[7] Mao, J., Q. Yu, and K. Aizawa, Noisy localization annotation refinement for object detection. IEICE Transactions on Information and Systems, 2021. 104(9): p. 1478-1485.

[8] Xu, Y., et al., Training robust object detectors from noisy category labels and imprecise bounding boxes. IEEE Transactions on Image Processing, 2021. 30: p. 5782-5792.

[9] Lin, T.-Y., et al. Microsoft coco: Common objects in context. In European conference on computer vision. 2014. Springer.

[10] Everingham, M., et al., The pascal visual object classes (voc) challenge. International journal of computer vision, 2010. 88(2): p. 303-338.

[11] Deng, J., et al. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. 2009. Ieee.

[12] Hodge V, Austin J. A survey of outlier detection methodologies[J]. Artificial intelligence review, 2004, 22(2): 85-126.

[13] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey[J]. ACM computing surveys (CSUR), 2009, 41(3): 1-58.

[14] Gamberger D, Lavrač N. Conditions for Occam's razor applicability and noise elimination[C]//European conference on machine learning. Springer, Berlin, Heidelberg, 1997: 108-123.

[15] Gamberger D, Lavrac N, Groselj C. Experiments with noise filtering in a medical domain[C]//ICML. 1999, 99: 143-151.

[16] Frénay B, Verleysen M. Classification in the presence of label noise: a survey[J]. IEEE transactions on neural networks and learning systems, 2013, 25(5): 845-869.

[17] Zhu, X. and X. Wu, Class noise vs. attribute noise: A quantitative study. Artificial intelligence review, 2004. 22(3): p. 177-210.

[18] Flatow, D. and D. Penner, On the robustness of convnets to training on noisy labels. Technical report, Stanford University, 2017.

[19] Nazari, Z., et al., Evaluation of class noise impact on performance of machine learning algorithms. IJCSNS Int. J. Comput. Sci. Netw. Secur, 2018. 18: p. 149.

[20] Xu, M., et al. Missing Labels in Object Detection. In CVPR Workshops. 2019.

[21] Li, Y. and X. Chao, Toward sustainability: trade-off between data quality and quantity in crop pest recognition. Frontiers in Plant Science, 2021. 12.

[22] Algan, G. and I. Ulusoy, Image classification with deep learning in the presence of noisy labels: A survey. Knowledge-Based Systems, 2021. 215: p. 106771.

[23] Tan, M. and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning. 2019. PMLR.

[24] Tan, M. and Q. Le. Efficientnetv2: Smaller models and faster training. In International Conference on Machine Learning. 2021. PMLR.

[25] Liu, C., et al., Robust Object Detection With Inaccurate Bounding Boxes. arXiv preprint arXiv:2207.09697, 2022.

[26] Muhammad, M.B. and M. Yeasin. Eigen-cam: Class activation map using principal components. in 2020 International Joint Conference on Neural Networks (IJCNN). 2020. IEEE.