*data*

MDPI

*Data Descriptor*

# LeLePhid: An Image Dataset for Aphid Detection and Infestation Severity on Lemon Leaves

**Jorge Parraga-Alava [1,*], Roberth Alcivar-Cevallos [1], Jéssica Morales Carrillo [2], Magdalena Castro [1], Shabely Avellán [1], Aaron Loor [2] and Fernando Mendoza [2]**

[1] Facultad de Ciencias Informáticas, Universidad Técnica de Manabí, Avenida Jose María Urbina, Portoviejo 130104, Ecuador; roberth.alcivar@utm.edu.ec (R.A.-C.); Magdalena.castro@utm.edu.ec (M.C.); Shabely.avellan@utm.edu.ec (S.A.)

[2] Carrera de Computación, Escuela Superior Politécnica Agropecuaria de Manabí, Sitio El Limón, Calceta 130250, Ecuador; jmorales@espam.edu.ec (J.M.C.); aaron.loor@espam.ec (A.L.); fernando.mendoza@espam.ec (F.M.)

* Correspondence: jorge.parraga@usach.cl; Tel.: +593-9-6731-7778

**Abstract:** Aphids are small insects that feed on plant sap, and they belong to a superfamily called *Aphoidea*. They are among the major pests causing damage to citrus crops in most parts of the world. Precise and automatic identification of aphids is needed to understand citrus pest dynamics and management. This article presents a dataset that contains 665 healthy and unhealthy lemon leaf images. The latter are leaves with the presence of aphids, and visible white spots characterize them. Moreover, each image includes a set of annotations that identify the leaf, its health state, and the infestation severity according to the percentage of the affected area on it. Images were collected manually in real-world conditions in a lemon plant field in Junín, Manabí, Ecuador, during the winter, by using a smartphone camera. The dataset is called LeLePhid: lemon (Le) leaf (Le) image dataset for aphid (Phid) detection and infestation severity. The data can facilitate evaluating models for image segmentation, detection, and classification problems related to plant disease recognition.

**Dataset:** https://doi.org/10.17632/tndhs2zng4

**Dataset License:** CC-BY

**Keywords:** plant disease recognition; image segmentation; image classification; aphid; *Aphoidea*; lemon



**Citation:** Parraga-Alava, J.; Alcivar-Cevallos, R.; Morales Carrillo, J.; Castro, M.; Avellán, S.; Loor, A.; Mendoza, F. LeLePhid: An Images Dataset for Aphids Detection and Infestation Severity on Lemons Leaf. *Data* **2021**, *6*, 51. https://doi.org/10.3390/data6050051

Academic Editors: Munish Kumar, R. K. Sharma and Ishwar Sethi

Received: 14 April 2021
Accepted: 12 May 2021
Published: 17 May 2021



## 1. Summary

The dataset, called LeLePhid in short, provides images of lemon leaves. This dataset contains 665 photos of the top and back of lemon tree leaves in which there are healthy and unhealthy leaves; these were collected manually in citrus crops from Junín, Ecuador, in winter, from December to May, when the weather is warm and rainy in this country. For the annotation process, it was carried out with the Labelbox© annotation tool, and to assign the severity of the infestation, three annotators manually inspected the image and set the grade of infestation severity according to [1] and the OIRSA method [2]. These data can be used for training, testing, and validation of computational models related to image segmentation and object detection in plant disease studies. At the same time, they can be helpful for researchers and professionals working on computer vision-based models for image classification and object detection using images of healthy leaves and leaves with the presence of aphids. The data annotations can be used to develop and improve the accuracy of lemon leaf aphid infestation severity and detection algorithms.

## 2. Data Description

The LeLePhid dataset provides lemon leaf images that can be used to develop and evaluate the performance of models of image segmentation, object detection, and classification problems related to plant diseases. The dataset contains imagery of the upper and back sides of leaves of lemon trees manually collected in citrus crops around Junín, Ecuador. On each image, the foreground leaf is identified, and its status is labeled, i.e., healthy and aphid[1] presence. The dataset also includes annotations to identify the infestation severity of the leaves affected by aphids. It can be used to design automatic aphid counting models because, as stated in [3], compared with manual counting[2], these models can calculate the percentage of the affected area through analyzing the image information. The released files for the so-named LeLePhid dataset are two folders: the raw data are available in the "Images" folder (665 images of lemon leaves) and pre-processed data are available in the "Annotation" folder (.json and .xlsx files). Samples of them are depicted in Figures 1 and 2. Figure 1 shows an example of the annotated images for segmentation purpose. In a green limited-area is identified the a lemon leaf. In purple areas the aphids presence. In Figure 2A, the class of the image is healthy meanwhile in Figure 2B the class is aphids, i.e., the leaf has presence of this insect. In addition, Tables 1 and 2 describe the levels or infestation severity on each lemon leaf available in the dataset. Finally, Figure 3 describes the distribution of images by health status and levels of infestation of aphids.

**Figure 1.** Annotation examples of a segmentation mask in the LeLePhid dataset.

[1] Aphids are tiny insects that feed by sucking sap from plants, and they can cause diseases.
[2] An adhesive board is placed on the plants, and researchers count the aphids on it.

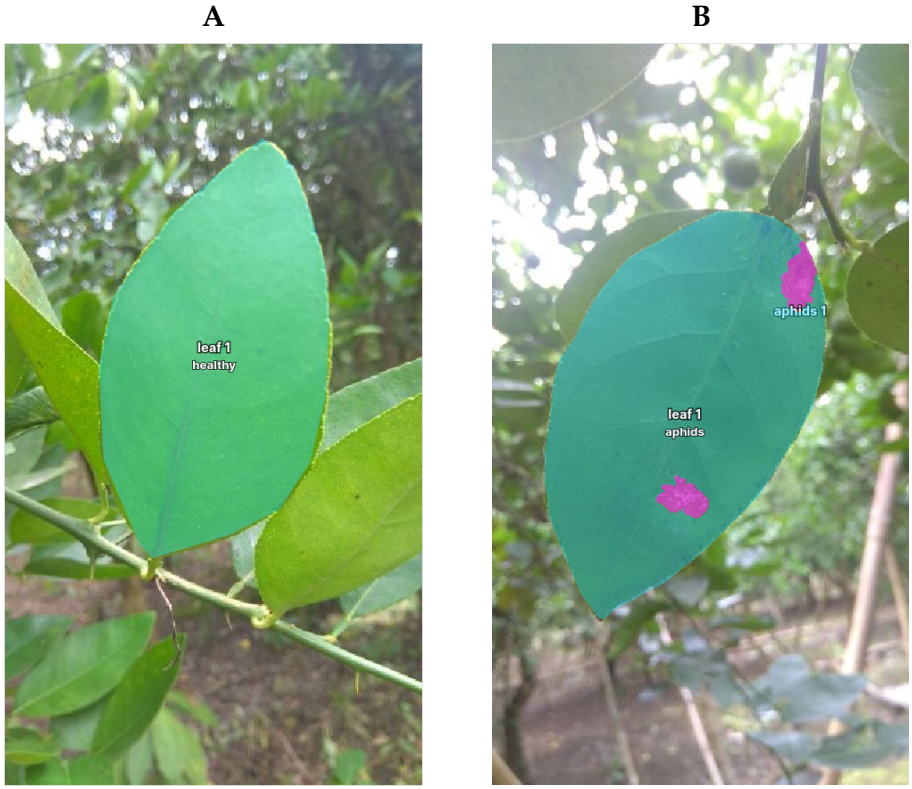


**Figure 2.** Annotation examples of lemon leaf image classification.

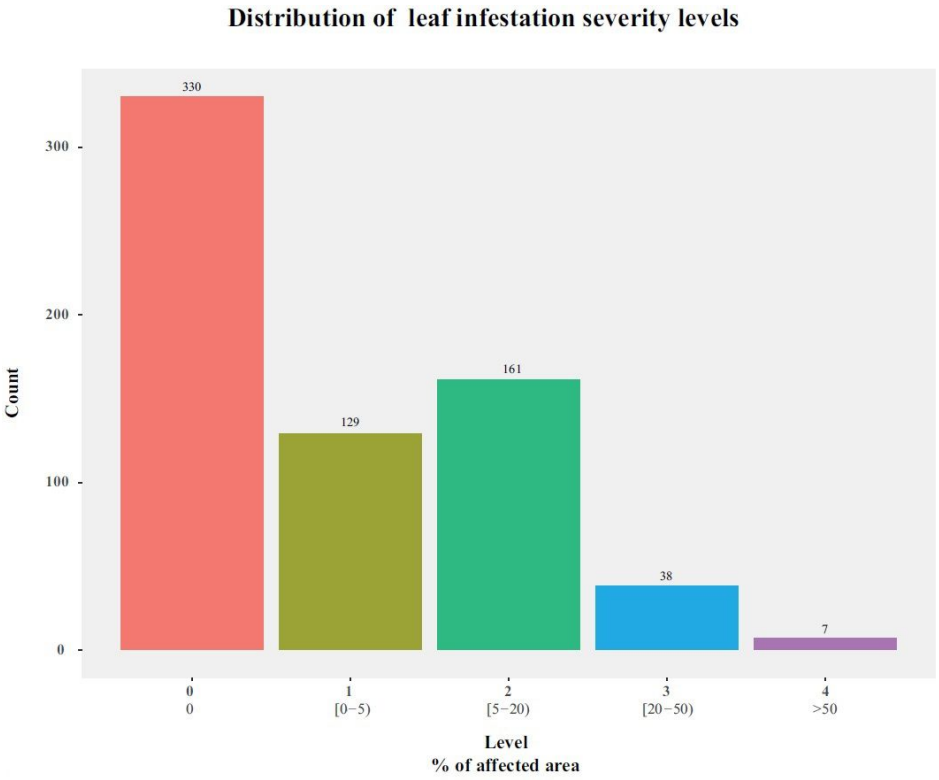


**Figure 3.** Distribution of images according to infestation severity levels.

## 3. Methods

LeLePhiD is designed to support computer vision research related to image processing with a particular focus on the detection and infestation severity of aphids on lemon leaves. The pipeline of the creation of this dataset is shown in Figure 4.

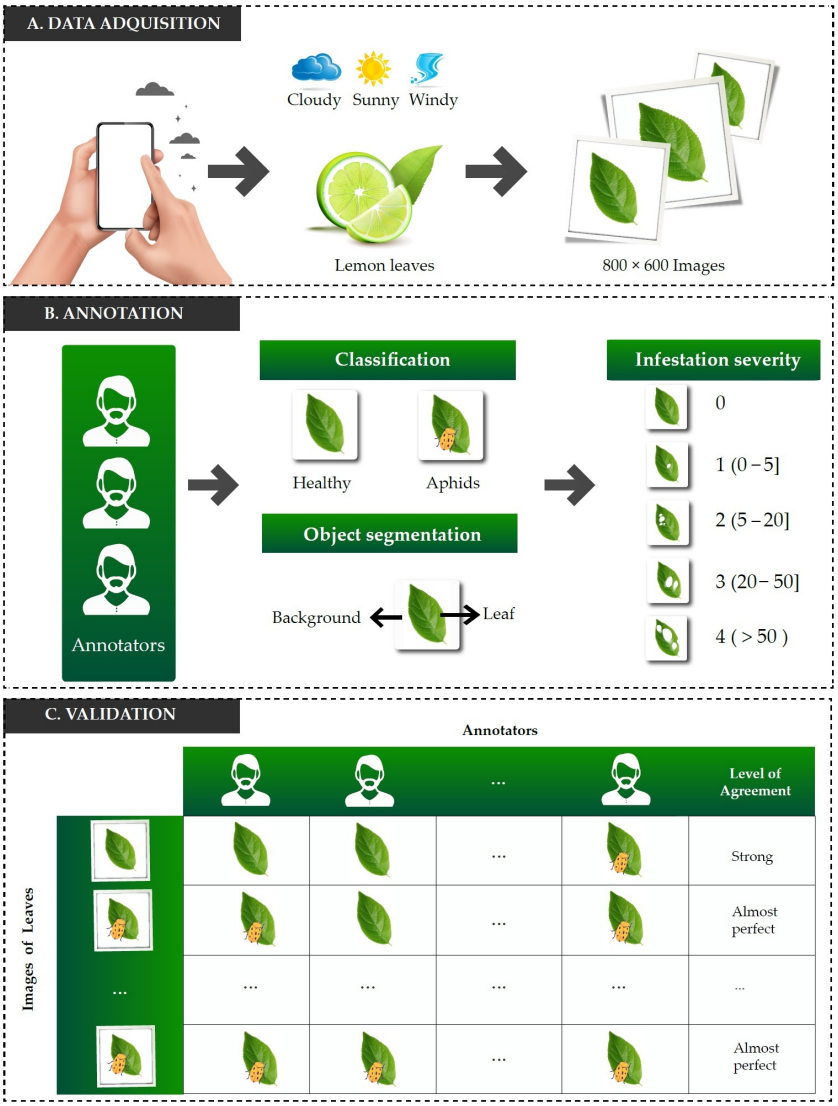


**Figure 4.** Pipeline of creation of LeLePhid dataset.

In Figure 4, we can see three steps, including the data acquisition, the incorporation of annotations, and the validation. In the following subsections, we detail each part of the process of creating this dataset.

### *3.1. Data Acquisition*

The lemon leaf images were manually acquired on a crop field in a rural area of Junín, Manabí, by using a 2 megapixel smartphone camera. Lemon images were captured following the procedure in [4] during cloudy, sunny, rainy, and windy days. The images were taken at a distance of 30–50 cm from the plant. The data capture process was performed in a time window of two weeks with different climatic conditions and background scenarios. We took 665 leaf images of the upper and back sides of healthy and unhealthy lemon plants. All images were rotated to a vertical position and resized to 800 × 600 pixels, keeping the aspect ratio. The process can be observed in Figure 4A.

### 3.2. Annotations

The annotation process was performed by using the Labelbox© annotation tool and can be observed in Figure 4B. In the object segmentation annotation, for each image, the foreground leaf is identified, and also, if the leaf is diseased, the area affected by aphids is marked (Figure 1). In the classification annotation, each image is labeled healthy or aphids according to the leaf health status (Figure 2). These annotations were assigned based on the comprehensive evaluation of the images of leaves according to the experience of the annotators.

Figure 1 shows an example of an annotation where the green-limited area identifies a lemon leaf and the magenta areas show the presence of aphids.

In Figure 2, the labeled lemon leaf images are shown. In Figure 2A, the image class is healthy; meanwhile, in Figure 2B, the class is aphids, i.e., the leaf has the presence of this insect. Note that only certain areas with white spots and texture correspond to aphids. Other spots related to other leaf conditions are not considered in this study.

To assign the infestation severity of each leaf, three annotators manually inspected the image and set the grade of infestation severity according to [1] and the OIRSA method [2]. The description of the levels of infestation severity grades of the affected area in lemon leaves can be observed in Table 1.

**Table 1.** Infestation severity scale of aphids in plants.

| Level | % Affected Area | Symptom |
|---|---|---|
| 0 | 0 | Healthy plant with no aphid presence. |
| 1 | [0–5) | Few aphids. Foliage with no yellowing symptoms. |
| 2 | [5–20) | Crinkling and curling of few leaves of the plant. |
| 3 | [20–50) | Crinkling and curling of leaves almost all over the plant. |
| 4 | >50 | Extreme curling, crinkling, and drying all over the plant. |

### 3.3. Validation

The validation process can be observed in Figure 4C. The consistency of the annotations was validated using the agreement between annotators. This was achieved by seeking matches in the category assigned to each image by the annotators. To quantify this, we used the kappa coefficient and the interpretation suggested by [5]. It can be simplified in Table 2 as follows:

**Table 2.** Interpretation of Cohen's kappa.

| Kappa | Level of Agreement | % of Data Reliability |
|---|---|---|
| 0–0.20 | None | 0–4% |
| 0.21–0.39 | Minimal | 4–15% |
| 0.40–0.59 | Weak | 15–35% |
| 0.60–0.79 | Moderate | 35–63% |
| 0.80–0.90 | Strong | 64–81% |
| Above 0.90 | Almost Perfect | 82–100% |

In Table 2, any kappa value below 0.60 indicates inadequate agreement among the annotators and that little confidence should be placed in the labeling process. Here, the percentage of data reliability corresponds to the squared kappa value. The final value of each label (level) was selected using a plurality strategy, i.e., when the matches are greater than 2. In cases of ties, the value was arbitrarily chosen in random order. The level of agreement obtained by our annotators was 91.0%, which means that the real percentage of affected area by aphids with LeLePhid is almost perfect.

Finally, the LeLePhid dataset contains 665 lemon leaf images distributed into 330 healthy leaves and 335 leaves with aphid presence. The latter are categorized according to leaf infestation severity and distributed as summarized in Figure 3.

In Figure 3, we include the image distribution according to the infestation severity levels. Note that there are 330 images with 0% of affected area, i.e., they correspond to healthy photos. The 335 images with aphid presence are divided into four levels. The first one has 129 leaf images with less than 5% of affected area (level 1). Most of the leaf images (161) are of level 2, i.e., they have between 5% and 20% of affected area. Further, there are 38 images with between 20% and 50% of infestation (level 3). Finally, the dataset contains seven leaf images with more than 50% of affected area (level 4).

## 4. User Notes

The data described in this paper are from a citrus crop near Junín, Ecuador (latitude −0.9277, longitude −80.2058). They were acquired using a smartphone camera. The identifications of the leaf, its state, and the area affected by aphids were individually incorporated as annotations over the image. The annotation is provided as a JSON file supported in any computer vision software. The possibilities of practical application are the following:

- The data can be used to train, test, and validate computational models related to image classification on plant disease studies. In this sense, we already have evidence from a previous work [6], where convolutional neural networks (CNNs) were used to board a binary classification problem related to lemon leaves with aphid presence. The quality of LeLePhid was evidenced by allowing the model to achieve average rates between 81% and 97% of correct aphid classification.
- The data can be helpful to researchers and professionals working on computer vision-based models for image segmentation and object detection using images of healthy leaves and leaves with aphid presence. Cases such as those discussed in [3,7] are examples of the potential that our dataset can offer from the point of view of continuous improvement of machine learning algorithms to address segmentation and identification problems related to plant diseases.
- The data can serve as a motivation to encourage further research into the agriculture sector and computer vision methods for citrus pest identification. Image annotation is the data labeling technique used to make the varied objects recognizable for computers. Our dataset includes image annotations of leaves and aphid-infected areas to make them recognizable or even understandable for computers. These annotations can be used to help the large-scale monitoring of the health of crops through, for instance, devices such as UAVs (unmanned aerial vehicles) or drones, where works in [8–11] have already demonstrated the benefits that can be obtained in the agricultural sector when devices such as drones are used in conjunction with computer vision.

Note that most of the images used by algorithms of the two first bullet points were captured in controlled environments, i.e., computer vision laboratories where the photos are treated artificially: constant backgrounds, homogeneous luminosity, and other conditions not usually occurring in lemon crops. Our dataset stands out from the others because the images were captured during cloudy, rainy, sunny, and windy days and considered scenarios with a variety of backgrounds in a typical lemon crop. This ensures that the algorithms learn from representative images of the type and complexity of real-world scenes.

**Author Contributions:** J.P.-A.: conceptualization, methodology, investigation, resources, data curation, writing—original draft, writing—review. R.A.-C.: data curation, investigation, writing—original draft, writing—review. J.M.C.: writing—review. M.C.: data curation. S.A.: data curation. A.L.: data curation. F.M.: data curation. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Data regarding images and annotations can be accessed at: repository name: *LeLePhid*; data identification number: DOI: 10.17632/tndhs2zng4; direct URL to data: https://data.mendeley.com/datasets/tndhs2zng4; accessed date: 13 January 2021.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Chen, T.; Zeng, R.; Guo, W.; Hou, X.; Lan, Y.; Zhang, L. Detection of Stress in Cotton (*Gossypium hirsutum* L.) Caused by Aphids Using Leaf Level Hyperspectral Measurements. *Sensors* **2018**, *18*, 2798. [CrossRef] [PubMed]
2. Virginio-Filho, E.; Astorga, C. *Prevención y Control de la Roya del Café. Manual de Buenas Prácticas para Técnicos y Facilitadores*; Centro Agronómico Tropical de Investigación y Enseñanza (CATIE): Turrialba, Costa Rica, 2015.
3. Suo, X.; Liu, Z.; Sun, L.; Wang, J.; Zhao, Y. Aphid Identification and Counting Based on Smartphone and Machine Vision. *J. Sens.* **2017**, *2017*, 3964376:1–3964376:7.
4. Parraga-Alava, J.; Cusme, K.; Loor, A.; Santander, E. RoCoLe: A robusta coffee leaf images dataset for evaluation of machine learning based methods in plant diseases recognition. *Data Brief* **2019**, *25*, 104414. [CrossRef] [PubMed]
5. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Medica* **2012**, *22*, 276–282. [CrossRef]
6. Parraga-Alava, J.; Alcivar-Cevallos, R.; Riascos, J.A.; Becerra, M.A. Aphids Detection on Lemons Leaf Image Using Convolutional Neural Networks. In *Systems and Information Sciences*; Botto-Tobar, M., Zamora, W., Larrea Plúa, J., Bazurto Roldan, J., Santamaría Philco, A., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 16–27.
7. Chen, J.; Fan, Y.; Wang, T.; Zhang, C.; Qiu, Z.; He, Y. Automatic Segmentation and Counting of Aphid Nymphs on Leaves Using Convolutional Neural Networks. *Agronomy* **2018**, *8*, 129. [CrossRef]
8. Bah, M.D.; Hafiane, A.; Canals, R. Deep Learning with Unsupervised Data Labeling for Weed Detection in Line Crops in UAV Images. *Remote Sens.* **2018**, *10*, 1690. [CrossRef]
9. Cãlina, J.; Cãlina, A.; Miluț, M.; Croitoru, A.; Stan, I.; Buzatu, C. Use of drones in cadastral works and precision works in silviculture and agriculture. *Rom. Agric. Res.* **2020**, *37*, 273–284.
10. Kitano, B.T.; Mendes, C.C.T.; Geus, A.R.; Oliveira, H.C.; Souza, J.R. Corn Plant Counting Using Deep Learning and UAV Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, 1–5. [CrossRef]
11. Gomez Selvaraj, M.; Vergara, A.; Montenegro, F.; Alonso Ruiz, H.; Safari, N.; Raymaekers, D.; Ocimati, W.; Ntamwira, J.; Tits, L.; Omondi, A.B.; et al. Detection of banana plants and their major diseases through aerial images and machine learning methods: A case study in DR Congo and Republic of Benin. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *169*, 110–124. [CrossRef]