Article

DiaMOS Plant: A Dataset for Diagnosis and Monitoring Plant Disease

Gianni Fenu [1], **Francesca Maridina Malloci** [2*]

[1] Department of Mathematics and Computer Science, University of Cagliari Via Ospedale 72, 09124 Cagliari, Italy; fenu@unica.it
[2] Department of Mathematics and Computer Science, University of Cagliari Via Ospedale 72, 09124 Cagliari, Italy; francescam.malloci@unica.it
* Correspondence: francescam.malloci@unica.it; fenu@unica.it

Version October 8, 2021 submitted to Agronomy

Abstract: The classification and recognition of foliar diseases is an increasingly developing field of research, where the concepts of machine and deep learning intervene to support agricultural stakeholders. Datasets are the fuel for the development of these technologies. In this paper, we release publicly available the field-dataset collected to diagnose and monitor plants symptoms, called DiaMOS Plant, consisting of 3505 images of pear fruit and leaves affected by four diseases. In addition, we perform a comparative analysis of existing literature datasets designed for the classification and recognition of leaf diseases, highlighting the main features that maximize the value and information content of the collected data. This study provides guidelines that will be useful to the research community on data set selection and construction.

Keywords: Plant Disease Prediction; Classification; Detection; Dataset; Survey; Machine learning; Deep Learning;

1. Introduction

Direct visual analysis of the leaves provides valuable information on plant health. Leaf symptoms are the first warning signs of many diseases, infections, parasites and deficiencies that occur during the development and life cycle of the plant. Biotic and abiotic stresses represent the main factors limiting agricultural productivity, such as to cause huge production losses.

An economic-environmental issue that is attracting increasing attention, becoming a hotspot in research [1], due to intensifying pressure from climate change and an estimated increase in world population of 70% by 2050 that will grow food demand [2]. A challenge that finds a solution in innovation and the development of sustainable cultivation practices that make efficient use of available resources.

The promotion of qualitatively and quantitatively sustainable actions is made possible by the adoption of recent information and communication technologies, the so-called ICT. The use of proximity sensors is driving the entry into the field of operational IT tools capable of assisting the farmer in cultivation practices. Mobile and robotic applications are the enabling solutions for the digital innovation process needed to safeguard the planet by assisting in monitoring and treatment operations. The integration of Artificial Intelligence [3] [4] in these systems is indispensable to support the operator in making informed and thoughtful decisions on the real state of the vigour of the plant. These tools are able to support stakeholders in both early prediction and diagnosis by recognizing symptoms visible to the naked eye. In the first task, the models are categorized into three categories [1]: (i) forecast model based on weather data; (ii) forecast models based on image processing; (iii) forecast models based on distinct types of data coming from various heterogeneous sources. The second task, diagnosis is mainly performed by processing RGB, multispectral or remote sensing images. In this context, Computer Vision [5] finds a relevant application, which by using appropriate networks trained on image samples, can detect, recognise and identify situations of crop risk and identify the various stages of fruit growth, useful for mechanical harvesting. Recent literature is addressing the problem with training single-output or multi-output convolutional neural networks [5], an approach known as Multitask learning.

The accuracy and reliability of integrated artificial intelligence systems is highly influenced by the representativeness and completeness of the dataset used in training the algorithm. The development of intelligent neural networks needs large quantities of data to be able to learn, from known examples, the essential knowledge to obtain a greater generalizability of the model. However, the realisation of a dataset, is not a simple and immediate task, due to the efforts and costs required that range from the acquisition, annotation and categorisation of the images, which often must be carried out by different professional figures expert in the sector. The availability of datasets in Digital Agriculture (DA) has become a well-known problem in the literature, slowing down scientific progress [6].

In recent years, several efforts have been made in data collection. Several datasets have been introduced. The best known in this field is PlantVillage [7], consisting of 54,000 images, portrayed on the ventral side of the leaf, on a homogeneous background. However,as observed by the literature [8] these configurations are not sufficiently representative for the objectives of the final application. The datasets created under controlled conditions, i.e. depicting the leaf on a homogeneous background, do not realistically reproduce the possible environmental conditions in which the model will operate.

In this context, the contribution of this paper is articulated on two levels. We introduce a new dataset in the literature for the diagnosis and monitoring of plant symptoms, called DiaMOS Plant. It is a dataset collected under realistic field conditions, composed of 3505 images depicting 4 leaf stresses and 3 stages of fruit development, such as fruit set, growth and ripening. We conduct a survey dedicated to public image datasets built for the classification and identification of leaf diseases. We focus on datasets released in open format on data sharing platforms. Therefore, we do not deal with datasets released under request to authors. The development and release of publicly available datasets has a twofold advantage. It allows researchers to save time and resources, and devote more effort to objective evaluation and comparison of algorithms. A research work was conducted for various tasks related to computer vision in the context of precision agriculture [9]. This survey seeks to cover the lack of a complete description for this particular sub-field. We believe that this survey would be a useful resource in guiding insightful selection of datasets for future research.

The rest of the paper is organised as follows. Section II describes the proposed DiaMOS Plant dataset and summarizes the characteristics of the publicly available image datasets. Section III provides a comparative analysis of the examined datasets. Section IV, provides some recommendations on requirements for future creation of datasets and a brief conclusion is drawn.

2. DiaMOS Plant dataset

In this section we describe in detail the proposed dataset.

Description. In this work, we introduce a field dataset to diagnose and monitor plants' symptoms called DiaMOS Plant, an extended dataset analyzed in [5]. DiaMOS Plant is a pilot dataset contains images of an entire growing season of pear tree, from February to July, in order to build a representative sample which, cover the main cultural aspects of this plant. The dataset is suitable to perform machine and deep learning methods in classification and detection tasks. A total of 3505 images were collected, including 499 fruit images and 3006 leaves images, respectively. The fruit is portrayed in the following 4 phases: fruit set, nut fruit, fruit growth, ripening. Similarly, biotic and abiotic stresses fall into 4 categories: leaf spot, leaf curl, slug damage, and healthy leaf. A detailed summary is provided in Tables 1, 2.

DiaMOS Plant Dataset	
Plant	Pear
Cultivar	Septoria Piricola
Data Source Location	Sardegna, Italy
Type of data	RGB Images
Annotation	csv, YOLO
ROI (Region of Interest) captured	leaf, fruit
Total size	3505 images (3006 leaves images + 499 fruit images)
Data Accessibility	Direct URL to data: https://doi.org/10.5281/zenodo.5557313
Application	The images are suitable for different machine and deep learning tasks such as images detection and classification.

Table 1. Dataset Descirption.

Leaves images	*Leaf Symptoms*	*Size*
	Healthy	43
	Spot	884
	Curl	54
	Slug	2025
	Severity Levels	*Size*
	0	43
	1	682
	2	1139
	3	699
	4	389

Table 2. DiaMOSP Plant is a collection of 3505 images of fruits and leaves. The table illustrates the distribution of classes belonging to the leaf images.

	Smartphone camera	DSRL camera
Image size	2976 X 3968	3456 X 5184
Model device	Honor 6X	Canon EOS 60D
Focal length	3,83 mm	50 mm
Focal ratio	f/2,2	f/4,5
Color space	RGB	RGB

Table 3. Acquisition device configurations.

Figure 1. On the first row, from left to the right, images of pear leaves captured under different light conditions: indirect sunlight, direct sunlight, strong sunlight reflection, distributed light. On the second row, images of pear fruit in different stages of growth.

The images belong to three trees have available from the same plot located in Italy. Pictures were gathered using different devices including a smartphone (Honor 6x) and DSRL camera (Canon EOS 60D), thus the images present two type of resolutions, 2976 X 3968 and 3456 X5184 respectively. Table 3 reports the set-up of each device. We employed two different devices because more people were involved in collecting data, and it was not feasible have the same devices. Furthermore, the different resolution increases the complexity of the dataset and represents an added value to it. The choice of using multiple devices is a widely used approach in this field of literature as it allows to provide heterogeneous and representative inputs to the models. In the real scenario, agricultural and non-agricultural operators have a smartphone that differs in different technical characteristics, including resolution.

The leaves were captured from the adaxial (upper) side of the leaf, in a real-life scenario where they were shot in various lighting (cloudy, sunny and windy days), angles, backgrounds (other plants and weeds) and noise conditions, at different times of the day throughout the entire growing season. This acquisition protocol has made it possible to obtain numerous advantages, such as: (i) capturing leaves under realistic lighting conditions that can be classified as: (a) indirect sunlight, (b) direct sunlight, (c) strong reflection (d) evenly distributed light (see Fig. 1); (ii) capturing the evolution of visual symptoms; (iii) capturing the fruit from the fruit set phase to the ripening phase.

The disease recognition process for dataset labeling was assisted by an expert. The dataset was annotated manually using the LabelImg software [1]. Each original image of the entire leaf is labeled with the predominant disease. For healthy, leaf spot and slug damage classes, a severity level is assigned, where each level is set according to the percent of affected leaf area. The stress severity was calculated identifying five classes expressed as no risk (0%), very low (1–5%), low (6–20%), medium (21–25%), and high (>50%) in a range from 0 to 4 (see Table 2. The annotated labels are released in a csv format, while the bounding boxes are released in YOLO format. The dataset is freely available for academic purposes from a repository at https://doi.org/10.5281/zenodo.5557313 where the folder has the following structure:

```
DiaMOS Plant
├── description
└── pear
    ├── annotation
    │   ├── csv
    │   └── YOLO
    ├── leaves
    │   ├── spot
    │   ├── curl
    │   ├── slug
    │   └── healthy
    └── fruits
```

- *Description:* it contains the data description;
- *Pear:* it contains the data related pear tree;
- *Annotation:* It contains the annotation files;
- *Leaves:* it contains the leaves images;

[1] Tzutalin. LabelImg. Git code (2015). https://github.com/tzutalin/labelImg

- *Fruits:* it contains the fruit images.

News of dataset updates will be posted on the following site https://francescamalloci.com/category/projects/, as we will plan to continue to extend the dataset with additional fruit plants.

Benchmark dataset. In this section we provide a benchmark dataset, with the aim of providing a baseline for the classification task. In this regard, we compared the performances of five well-known convolutional neural network architectures, such as VGG19, ResNet50, InceptionV3, MobileNetV2, EfficientNetB0, as they are widely adopted in different classification tasks and have shown good generalization skills in the literature under review.

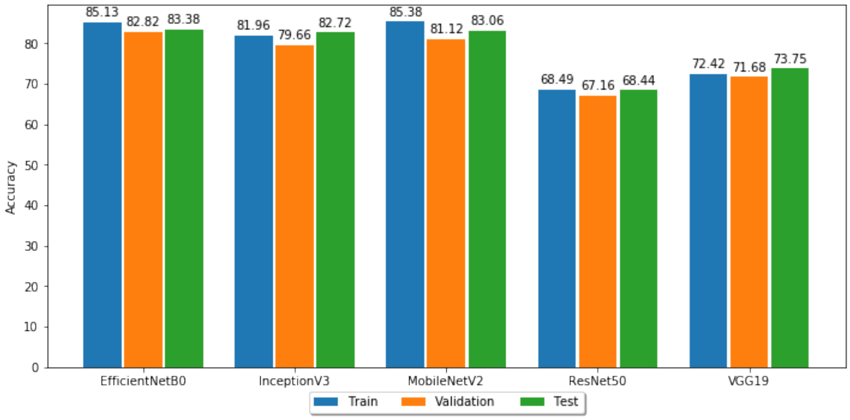
The experiment described here was conducted with the LeafBox toolbox developed and released in an open format, more purely for educational purposes and intended to facilitate the reproduction of our results and further research in this direction. It can be reached at the following link: https://github.com/mallociFrancesca/leaf-disease-toolbox.git. The experimental framework written in Python language exploits the Keras deep learning 2.4.3 library based on TensorFlow 2.2.1 environment, executed on a server machine with a 3.000GHz Intel® Xeon® Gold, and 64 Gb of memory [5], .

The classification task involved four ground truths, such as "healthy", "slug", "curl", "spot". The dataset was divided into training, validation, and test datasets with a ratio of 7:2:1, respectively. To preserve the percentage of samples for each class, the dataset is split using the ShuffleSplit strategy provided by scikit-learn 0.23.2 library. All images were resized to 224x224x3. In the training phase, to better manage the unbalance of the classes and minimize overfitting situations, the augmentation technique was applied, including horizontal and vertical mirroring, rotation, and color variation. To avoid a long training time, the transfer learning method is applied. The training was performed by adapting CNN networks trained using ImageNet dataset [10], with a cross-entropy function. Furthermore, we monitored the model's validation loss to reduce the learning rate when it has stopped improving, to get out the Plateau phenomenon. A learning-rate of 2e-5, and a Momentum of 0.9, were set. The settings were identified by carrying out various tests, and on the basis of the results, those were chosen that allow to obtain models that are more robust and less affected by overfitting problems. The test was repeated twice, to record the model's performance with the RMSprop optimizer and the Adam optimizer.

Figure 2 and Table 4 report training, validation and test accuracy obtained with RMSprop optimizer; while Figure 3 and Table 5 report the results achieved with the Adam optimizer.

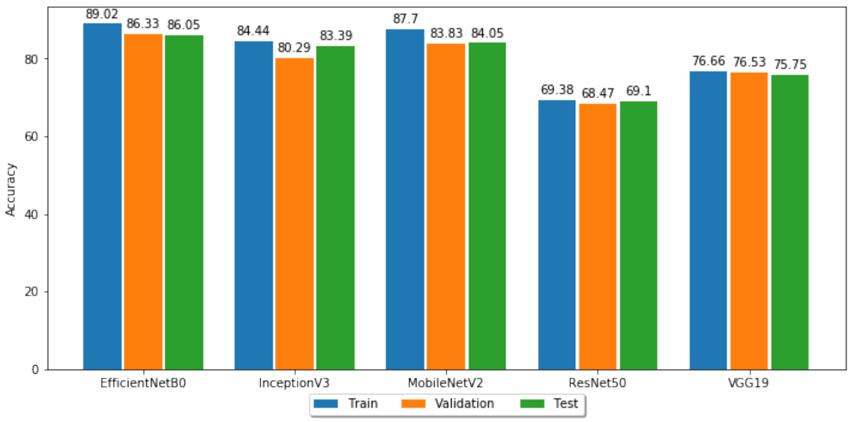
Comparing Tables 4, and 5 we observe similar performances for both optimizers, but there is a slight improvement with the Adam optimizer. However, this improvement is at the expense of the robustness of the results. Indeed, comparing the accuracy obtained in the three data sets, there is a more marked gap in the latter.

In general, it can be seen that the three networks EfficientNetB0, InceptionV3, and MobileNetV2 have a better generalization capacity than the VGG19 and ResNet50 networks. In fact, with reference to Table 4, EfficientNetB0, InceptionV3 and MobileNetV2 obtained an accuracy for the test set of 83.38 %, 82.72 %, 83.06 % respectively, while ResNet50 of 56.67 %, and VGG19 of 71.76 %. Comparing the scores recorded between the training, validation and test set, it is not excluded that the models may suffer from a slight overfitting bias. All things being equal, MobileNetV2 tends to converge faster. In Figure 5 and Table 5, the Precision, the Recall and the F1-score obtained in the test set are reported. Also in this case the f1-score ratio does not show notable differences in performance, reporting a high value for EfficientNetB0, InceptionV3, and MobileNetV2.



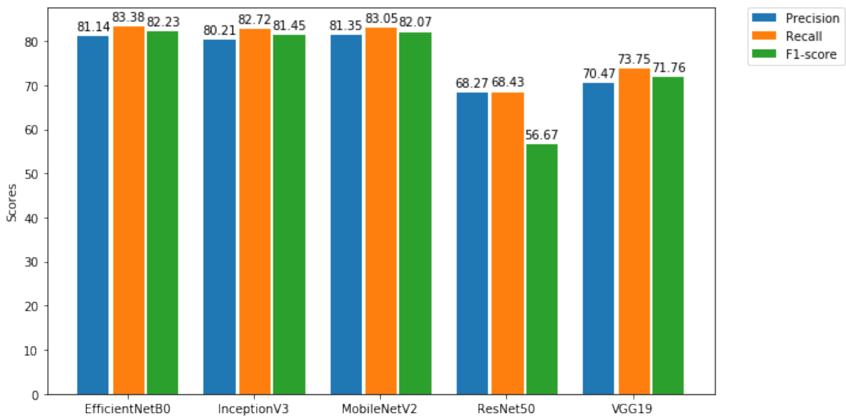
	RMSprop		
CNN	**Train Acc(%)**	**Validation Acc (%)**	**Test Acc (%)**
EfficientNetB0	81.13	82.82	83.38
InceptionV3	81.96	79.66	82.72
MobileNetV2	**85.38**	81.12	83.06
ResNet50	68.49	67.16	68.44
VGG19	72.42	71.68	73.75

Figure 2 & Table 4. Accuracy obtained with RMSprop optimizer respectively in the training set, validation set and test set in the task of classifying the "healthy", "slug", "curl", "spot" classes.



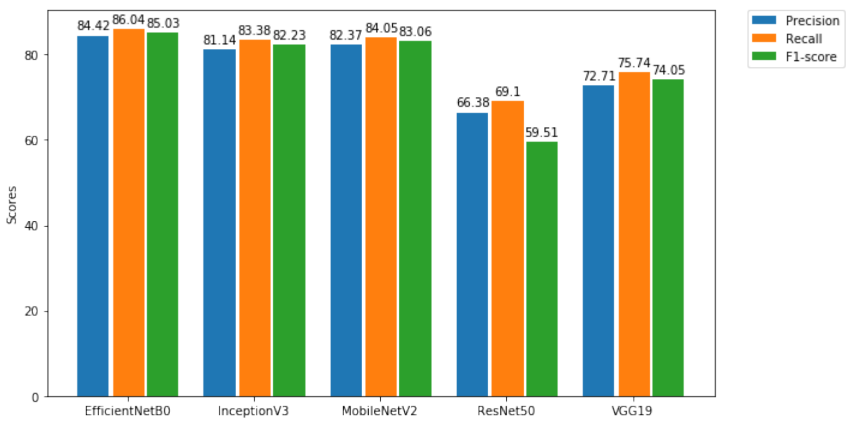
	Adam		
CNN	**Train Acc(%)**	**Validation Acc (%)**	**Test Acc (%)**
EfficientNetB0	**89.02**	86.33	86.05
InceptionV3	84.44	80.29	83.39
MobileNetV2	87.70	83.83	84.05
ResNet50	68.38	68.47	69.10
VGG19	76.66	76.53	75.75

Figure 3 & Table 5. Accuracy obtained with Adam optimizer respectively in the training set, validation set and test set in the task of classifying the "healthy", "slug", "curl", "spot" classes.



CNN	*RMSprop* Precision (%)	Recall (%)	F1-score (%)
EfficientNetB0	81.14	83.38	82.23
InceptionV3	80.21	82.72	81.45
MobileNetV2	81.35	83.05	82.07
ResNet50	68.27	68.43	56.67
VGG19	70.47	73.75	71.76

Figure 4 & Table 6. Precision, Recall, and F1-score reported with RMSprop optimizer on test set in the task of classifying the "healthy", "slug", "curl", "spot" classes.



CNN	*Adam* Precision (%)	Recall (%)	F1-score (%)
EfficientNetB0	84.42	86.04	85.03
InceptionV3	81.14	83.38	82.23
MobileNetV2	82.37	84.05	83.06
ResNet50	66.38	69.10	59.51
VGG19	72.71	75.74	74.05

Figure 5 & Table 7. Precision, Recall, and F1-score reported with Adam optimizer on test set in the task of classifying the "healthy", "slug", "curl", "spot" classes.

3. Open-Dataset for plant disease classification and detection

In this section we provide a brief description of the datasets presents in the literature.

3.1. RoCoLe dataset

RoCoLe is the acronymous of Robusta Coffee Leaf images dataset [11], containing 1560 leaf pictures divided into six classes: healthy, red spider mite presence, rust level 1, rust level 2, rust level 3 and rust level 4. The photos were captured from the adaxial (upper) and abaxial (lower) leaf side, under a natural uncontrolled environment, using a smartphone camera at a working distance of 200 and 300 mm without zoom. In addition, the dataset includes annotations regarding segmentation object, processed with the web-tool called Labelbox.

3.2. BRACOL dataset

BRACOL is a brazilian arabica coffee leaf images dataset to identification and quantification of coffee diseases and pests [12]. it contains 1747 images of arabica coffee leaves affected by the following biotic stresses: leaf miner, leaf rust, brown leaf spot, and cercospora leaf spot. The images were collected at different times of the year in Santa Maria of Marechal Floreano in the mountains regions of the state of Espirito Santo, Brazil. Obtained using five different smartphones the leaves were depicted from the abaxial (lower) side under partially controlled conditions and placed on a white background. The acquisition of the images was done without much criterion to make the dataset more heterogeneous. The process of biotic stresses recognition for dataset labeling was assisted by an expert.

3.3. Rice Leaf Disease dataset

The Rice Leaf dataset [13] consist of 120 images collected from a village called Shertha near Gandhinagar, Gujarat, India, captured with a white background using a Nikon D90 digital SRL camera with 12.3 megapixels in November 2015. The authors collected leaves having varying degree of disease spread, where all images have a resolution of 2848 x 4288 pixels.

3.4. Plant Pathology dataset

The Plant Pathology dataset [14] is a collection of 3651 RGB images of multiple apple foliar disease symptoms captured during the 2019 growing season from commercially grown cultivars in an unsprayed apple orchard at Cornell AgriTech (Geneva, New York, USA). Of the 3651 RGB images, there are 1200 of apple scab, 1399 of cedar apple rust, 187 of complex disease symptoms (i.e., more than one disease on the same leaf), and 865 of healthy leaves. Photos were taken using a Canon Rebel T5i DSLR and smartphones under various illumination, angle, surface, and noise conditions, directly from the field. The dataset was manually annotated into three classes: cedar apple rust, apple scab, multiple diseases, and healthy leaves. An expert plant pathologist confirmed the annotations.

3.5. Citrus dataset

The Citrus dataset [15] contain 759 images of healthy and unhealthy citrus fruits and leaves, manually acquired using a DSLR with the help of a domain expert. The infected images are classified into 4 different diseases of citrus fruits and leaves separately. The diseases present in the datasets are black spot, canker, scab, greening, and melanose. All images are resized to the dimension of 256*256 with 72 dpi resolution. The fruit images were collected directly from the plant, while leaves images were acquired under laboratory condition, with an homogeneous gray background.

3.6. APDA dataset

The APDA dataset [16] collected by Tea Research Institute, Mansehra, contains 40 images, divided into healthly and unhealthily. The diseased subset contains samples of two types of diseases:

anthracnose and black spots. Acquired with a Nikon camera D90, the leaves are depicted in indoor lighting, maintaining a constant distance of the object from the lens of approximately 9-12 inches.

3.7. PlantVillage dataset

The Plant Village is an image-based dataset of 54,309 samples in which foliar diseases are portrayed on the ventral side of the leaf, on a homogeneous background (black or gray). For each leaf, the authors took 4-7 images with a standard point and shoot camera Sony DSC - Rx100/13 with 20.2 megapixels, using the automatic mode. The images span 14 crop species: Apple, Blueberry, Cherry, Corn, Grape, Orange, Peach, Bell Pepper, Potato, Raspberry, Soybean, Squash, Strawberry, Tomato. In containes images of 17 fungal diseases, 4 bacterial diseases, 2 mold (oomycete) diseases, 2 viral disease, and 1 disease caused by a mite. 12 crop species also have images of healthy leaves that are not visibly affected by a disease.

4. Comparative Analysis

In this section we provide a comparative analysis of the examined datasets, including the proposed DiaMOS Plant dataset, organized into three sections: (i) dataset acquisition; (ii) symptoms and diseases; (iii) technical dataset settings. A summary scheme is shown in the Table 8.

4.1. Dataset Acquisistion

The place and mode of dataset acquisition influences how the algorithms learn and make predictions. The 62% of the datasets were collected under controlled conditions, using a mobile phone camera or DSRL camera. The remainder acquired the images directly in the field. The acquisition protocol followed by the laboratory datasets, in some studies was not characterised by certain criteria, in others it kept constant both the distance of the object of interest from the camera and the lighting conditions, portraying the leaf in the centre of the frame on a homogeneous background, mainly white.

With regard to the field datasets, the common goal was to maximise variability by adopting different techniques. Several acquisition tools were used. The leaf portrayed directly on the plant was acquired several times with different angles and illumination scenarios. The majority of cases, portrayed the leaf on the upper side, also called adaxial. Two exceptions are represented by BRACOL and RoCole, where RoCole portrayed both sides of the leaf (abaxial and adaxial) while BRACOL only portrayed the abaxial.

4.2. Symptoms and Diseases

In the plant world, there are many different stressful events that can give rise to the same or very similar visual symptoms. These events can also overlap and follow each other, making it even more complicated to arrive at an accurate and reliable diagnosis of the plant's condition [1]. Some researchers have taken into account the temporal variability in the evolution of the symptom from the first to the last stage. During a growing season, symptoms show different morphology, texture and colouration depending on the extent of the damage. For this purpose, DiaMOS Plant collected images at different times of the day for an entire growing season. This approach was also followed for the Plant Pathology dataset, which further enriched the dataset by annotating the presence of several diseases on the same leaf surface. Finally, DiaMOS Plant, BRACOL and RoCole labelled four levels of severity, useful to train models able to recognise the disease at different stages.

4.3. Technical Dataset Settings

Having a large dataset greatly affects the performance of machine and deep learning models. The datasets in this field are all small-scale datasets in terms of image number. Figure 6 shows the graphical distribution of the examined datasets according to size. PlantVillage is a large-scale dataset. However, certain classes contain few instances. As shown in Table 8, the RGB format was adopted by all studies,

Table 8. Details of datasets examined.

Dataset	DiaMOSPlant [5]	BRACOL [12]	RoCoLe [11]	Plant Pathology [14]	Rice Leaf Diseases [13]	Citrus [15]	APDA [16]	PlantVillage [7]
Plant / Crop	Pear	Coffee	Coffee	Apple	Rice	Citrus	Rose	Multiple
Dataset size	3505 (3006 leaf images + 499 fruit images)	4407	1560	3651	120	759 (609 leaf images + 150 fruit images)	40	54.309
n° symptoms	4	4	2	3	3	5	2	26
Acquisition device	Smarthphone e DSRL	Smartphone	Smartphone	DSLR Camera, Smartphone	DSLR camera	DSLR camera	Smartphone	Smartphone
Color	*RGB*	RGB	RGB	RGB	RGB	RGB	RGB	RGB
Image resolution	Multiple	2048x1024	Multiple	2048x1365	2848x4288	256 x 256	N.d.	Multiple
Annotation	Polygon, Label	Polygon, Label	Polygon, Label	Label	Label	Label	Label	Label
Annotation format	csv, YOLO	csv	csv, COCO, JSON, Pascal VOC	csv	folder structure	folder structure	N.d.	folder structure
Data sharing platform	Zenodo	GitHub	Mendeley Data	Kaggle	UCI Machine Learning Repository	Mendeley Data	MathWorks	Github
Acquisition place	field	laboratory	field	field	laboratory	laboratory	laboratory	laboratory
Side of the leaf	adaxial	abaxial	adaxial, abaxial	adaxial	adaxial	adaxial	adaxial	adaxial
Object of interest	Fruit, leaf	leaf	leaf	leaf	leaf	Fruit, leaf	leaf	leaf

and the acquisition approach involved the camera of a smartphone or DSRL. No datasets made use of drones. The acquired images can be used for the classification task, as they are appropriately annotated with labels. DiaMOS Plant, RoCole and BRACOL also feature bounding-box annotation, which allows the datasets to be used for the detection task right from the start. The most commonly used annotation format is csv. Finally, the data sharing methods were different. The prevailing methodology used external services. According to Lu and Young [9], this good practice allows to guarantee data availability over time.

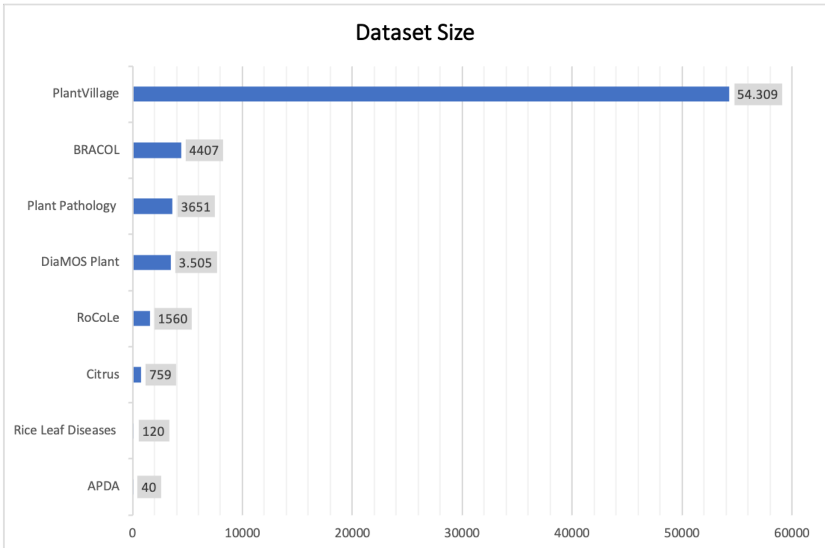


Figure 6. Graphical size distribution of the examined datasets.

Dataset	On-line Repository
DiaMOSPlant	https://doi.org/10.5281/zenodo.5557313
BRACOL [12]	https://data.mendeley.com/datasets/yy2k5y8mxg/1
RoCoLe [11]	https://data.mendeley.com/datasets/c5yvn32dzg/2
Plant Pathology [14]	https://www.kaggle.com/c/plant-pathology-2020-fgvc7
Rice Leaf Diseases [13]	https://archive.ics.uci.edu/ml/datasets/Rice+Leaf+Diseases
Citrus [15]	https://data.mendeley.com/datasets/3f83gxmv57/2
APDA [16]	https://it.mathworks.com/matlabcentral/fileexchange/55098
PlantVillage [7]	https://github.com/spMohanty/PlantVillage-Dataset

Table 9. Public image datasets with the related on-line repository.

5. Discussion

This analysis suggests that the most widely adopted image acquisition set up in the state-of-the-art is based on collected data under controlled, laboratory conditions. The analysis of current datasets have revealed some limitations including size, rappresentativeness, completness.

- *Dataset size:* the most limitations of current dataset is the small number of disease classes and samples size. Even our proposed dataset DiaMOS Plant, contains few samples for "healthy" class. Inevitably, a strong imbalance of classes leads to the model not generalising well in practical applications. This confirms and demonstrates, in agreement with Lu and Young [9], although the need for larger datasets is recognised, this task is challenging due to the manual effort and cost required, which in some cases is further exacerbated as very few occurrences in the field can occur for some classes. A technical problem that can be mitigated by data augmentation, transfer learning, and fine tuning techniques;
- *Representativeness:* The most widely adopted acquisition protocol is based on data collection under controlled, laboratory conditions. The representativeness of the dataset is limited by two

factors: place of acquisition, mode of acquisition. Controlled conditions are not able to reflect the spectrum of variability detectable in the field. As demonstrated by study in [17], algorithms tend to achieve near-perfect accuracy when trained on laboratory datasets, but performance degrades significantly when trained on field datasets. In addition, few datasets took into account the evolution of symptoms during an entire growing season. More efforts should focus on capturing symptoms at an early stage of emergency. In fact, at these stages digital aids are essential to take timely action to stop the disease proliferation.
- *Completness:* Strong *et al.* [18] define completeness as "the level of breadth, depth, and appropriateness of a datum according to its purpose". Although some datasets are well constructed, in some cases we found a lack of completeness in providing ground truth labels. The annotation of multiple symptoms present in the leaf maximises and completes the informative capacity of the data. Similarly, the presence of bounding-boxes and segmentation masks would extend their usability.
- *Performance Baseline:* The availability of a performance baseline can help the development and validation of new methods that can be applied.

Based on the limitations identified above, we provide some recommendations on creating future dataset. The number of sample and variety of diseases needs to be increased so that a learning algorithm may generalize on the problem domain. Algorithms are destined for inclusion in field applications, which can be categorized in:

- Disease recognition mobile applications;
- Robotic applications that recognize and identify the disease and spray chemical or natural inputs based on the extent of the damage.

To maximize the information content that the data can express, the completeness and representativeness of the samples, we suggest portraying the leaf using different configurations such as:

- Defer the angle, focus, position of the leaf in individual frames;
- Portrays the disease for an entire growing season, identifying different levels of severity;
- Collect the samples at different times of the day, that is with different climatic conditions (sunny, cloudy, direct light).

Finally, the dataset should be published on data sharing platforms, which allow the integrity and availability of data to be preserved over time [9].

6. Conclusion

In this paper, we released an open-dataset in the literature, called DiaMOS Plant, a self-collected dataset in the field, consisting of 3505 images, depicting 4 leaf diseases with 4 level of severity and 4 fruit stages, reachable at the following link https://doi.org/10.5281/zenodo.5557313. Simultaneously with the release of the dataset, we provided a performance baseline, and we reviewed the datasets present in the literature built for the classification and recognition of leaf diseases. The analysis conducted has highlighted the good practices for the construction of field data sets, which impact the information content that the data can express, as functional to its ability to describe the environment from which it was drawn or observed. Factors that were taken into consideration when constructing the proposed dataset. In this regard, for future works we plan to expand the released dataset, to enrich its representativeness and completeness, limited by the small number of samples for the "healthy" and "curled" class.

Author Contributions: All authors equally contributed to this research.

Acknowledgments: Francesca Maridina Malloci gratefully acknowledges Department of Mathematics and Computer Science of the University of Cagliari for the financial support of her PhD scholarship. We would like to thank the reviewers, whose valuable feedback, suggestions and comments increased significantly the overall quality of this study.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
DA	Digital Agriculture
ICT	Information and Communication Technologies

References

1. Fenu, G.; Malloci, F.M. Forecasting plant and crop disease: an explorative study on current algorithms. *Big Data and Cognitive Computing* **2021**, *5*, 2.
2. Food and Agriculture Organization of the United Nations. *The state of the world's land and water resources for food and agriculture: Managing systems at risk*; Earthscan, 2011.
3. Fenu, G.; Malloci, F.M. An application of machine learning technique in forecasting crop disease. Proceedings of the 2019 3rd International Conference on Big Data Research, 2019, pp. 76–82.
4. Fenu, G.; Malloci, F.M. Artificial intelligence technique in crop disease forecasting: a case study on potato late blight prediction. International Conference on Intelligent Decision Technologies. Springer, 2020, pp. 79–89.
5. Fenu, G.; Malloci, F.M. Using Multioutput Learning to Diagnose Plant Disease and Stress Severity. *Complexity* **2021**, *2021*.
6. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Computers and electronics in agriculture* **2018**, *147*, 70–90.
7. Hughes, D.; Salathé, M.; others. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060* **2015**.
8. Barbedo, J.G.A. Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering* **2019**, *180*, 96–107.
9. Lu, Y.; Young, S. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture* **2020**, *178*, 105760.
10. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009. doi:10.1109/cvprw.2009.5206848.
11. 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 in Brief* **2019**, *25*, 104414. doi:10.1016/j.dib.2019.104414.
12. Krohling, R.; Esgario, J.; Ventura, J.A. BRACOL - A Brazilian Arabica Coffee Leaf images dataset to identification and quantification of coffee diseases and pests. *Mendeley Data* **2019**, *V1*. doi:10.17632/yy2k5y8mxg.1.
13. Prajapati, H.B.; Shah, J.P.; Dabhi, V.K. Detection and classification of rice plant diseases. *Intelligent Decision Technologies* **2017**, *11*, 357–373.
14. Thapa, R.; Zhang, K.; Snavely, N.; Belongie, S.; Khan, A. The Plant Pathology Challenge 2020 data set to classify foliar disease of apples. *Applications in Plant Sciences* **2020**, *8*. doi:10.1002/aps3.11390.
15. Rauf, H.T.; Saleem, B.A.; Lali, M.I.U.; Khan, M.A.; Sharif, M.; Bukhari, S.A.C. A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning. *Data in brief* **2019**, *26*, 104340.
16. Akhtar, A.; Khanum, A.; Khan, S.A.; Shaukat, A. Automated plant disease analysis (APDA): performance comparison of machine learning techniques. 2013 11th International Conference on Frontiers of Information Technology. IEEE, 2013, pp. 60–65.

17. Fenu, G.; Malloci, F.M. Evaluating impacts between laboratory and field-collected datasets for plant disease classification. *Multimedia and Tools for Appplications* **2021**.
18. Strong, D.M.; Lee, Y.W.; Wang, R.Y. Data quality in context. *Communications of the ACM* **1997**, *40*, 103–110.