Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Plant disease identification from individual lesions and spots using deep learning

Jayme Garcia Arnal Barbedo

Embrapa Agricultural Informatics, Av. André Tosello, 209, C.P. 6041, Campinas, SP, 13083-886, Brazil

ARTICLE INFO

Article history:
Received 6 August 2018
Received in revised form
14 January 2019
Accepted 4 February 2019

Keywords:
Image processing
Image classification
Deep neural nets
Transfer learning
Disease classification

Deep learning is quickly becoming the standard technique for image classification. The main problem facing the automatic identification of plant diseases using this strategy is the lack of image databases capable of representing the wide variety of conditions and symptom characteristics found in practice. Data augmentation techniques decrease the impact of this problem, but those cannot reproduce most of the practical diversity. This paper explores the use of individual lesions and spots for the task, rather than considering the entire leaf. Since each region has its own characteristics, the variability of the data is increased without the need for additional images. This also allows the identification of multiple diseases affecting the same leaf. On the other hand, suitable symptom segmentation still needs to be done manually, preventing full automation. The accuracies obtained using this approach were, in average, 12% higher than those achieved using the original images. Additionally, no crop had accuracies below 75%, even when as many as 10 diseases were considered. Although the database does not cover the entire range of practical possibilities, these results indicate that, as long as enough data is available, deep learning techniques are effective for plant disease detection and recognition.

1. Introduction

Tools for automatic plant disease recognition have the potential to become a valuable source of information to aid decision making in farms (Barbedo, 2013). This is especially true in places where specialised technical support is not easily accessed and in large properties where continuous on-site monitoring is impractical. However, there are still many challenges lacking suitable solutions (Barbedo, 2016). Deep learning techniques, particularly Convolutional Neural Networks (CNN), are quickly becoming the preferred method to overcome some of those challenges (Barbedo, 2018a). Although very good results have been reported in the literature (Table 1), investigations so far have used image databases with limited diversity. In particular, many studies have employed the PlantVillage database (Amara, Bouaziz, & Algergawy, 2017; Brahimi, Boukhalfa, & Moussaoui, 2017; Cruz, Luvisi, Bellis, & Ampatzidis, 2017; Ferentinos, 2018; Mohanty, Hughes, & Salathé, 2016), which contains a substantial proportion of images with homogeneous backgrounds, especially in its early versions (Hughes & Salathé, 2015; Mohanty et al., 2016). Thus, only a limited subset of the entire range of possibilities is being considered both for training and testing the algorithms. The effects of those data constraints was illustrated by Mohanty et al. (2016), who observed a quick drop in accuracy when the model trained

E-mail address: jayme.barbedo@embrapa.br.
https://doi.org/10.1016/j.biosystemseng.2019.02.002
1537-5110/

Table 1 – Studies employing deep learning for plant disease recognition.

Reference	Network	Dataset	Accuracy
Amara et al. (2017)	CNN (LeNet architecture)	PlantVillage	92%–99%
Brahimi et al. (2017)	CNN (AlexNet, GoogLeNet)	PlantVillage	99%
Cruz et al. (2017)	CNN (Modified LeNet)	Olive tree images (own)	99%
DeChant et al. (2017)	CNN (Pipeline)	Corn images (own)	97%
Ferentinos (2018)	CNN (Several)	PlantVillage	99%
Fuentes, Yoon, Kim, and Park (2017)	CNN (Several)	Tomato images (own)	83%
Liu, Zhang, He, and Li (2018)	CNN (AlexNet)	Apple images (own)	98%
Lu, Yi, Zeng, Liu, and Zhang (2017)	CNN (AlexNet inspired)	Rice images (own)	95%
Mohanty et al. (2016)	CNN (AlexNet, GoogLeNet)	PlantVillage	99%
Oppenheim and Shani (2017)	CNN (VGG)	Potato images (own)	96%

using the PlantVillage database was applied to images collected online. Despite their limitations, these previous investigations have successfully demonstrated the potential of deep learning techniques.

Now that the ability of deep neural networks for plant disease recognition has been proven, CNN capabilities should be stressed by applying more realistic and varied image datasets. The dataset used in this work was built having diversity in mind. Prior to the subdivision into individual lesions and spots, it included 1567 images representing 79 diseases affecting 14 plant species. More importantly, images were captured under a wide diversity of conditions and included an extensive variety of symptom characteristics (Section 2). However, while the resulting database was representative in qualitative terms, it was quantitatively lacking. Many specific conditions had only a very small number of images associated. Even with the use of augmentation techniques, which artificially increase diversity for better generalization (Liu et al., 2018), such a number is not enough for proper training of deep neural networks (Barbedo, 2018b). Thus, the original images were segmented into individual lesions and spots, increasing the number of images to 46,409. Two additional benefits came from this procedure: a) since conditions may vary within the same leaf, data diversity was also increased; b) with single symptoms being considered, it was now possible to identify multiple diseases affecting the same leaf.

It is worth noting that while the dataset used in this work is more diverse than those used in previous studies, it is far from containing the entire diversity that can be found in practice. This is because capturing images in the field and correctly labelling them is a difficult, expensive and time consuming task. Thus, building a truly comprehensive database would require resources that are beyond the capabilities of most institutions. Some initiatives are using the concepts of social networks to overcome this limitation (Barbedo, 2018b), but the annotation process has some limitations that may lead to unreliable labels (Barbedo, 2018a). More representative image databases could also be achieved if different research groups made their own datasets available. As a step towards this goal, the dataset used in this work is being made freely available for research purposes in the address https://www.digipathos-rep.cnptia.embrapa.br/.

This study was carried out using a pretrained CNN using the GoogLeNet architecture. The accuracies obtained for each crop varied from 75% to 100%. Such a variation was caused by differences in the number of images, number of diseases, variety of conditions and, consequently, difficulty level. The overall accuracy using individual lesions and spots was 94%, higher than those obtained using the original images with (82%) and without (82%) manual background removal.

2. Materials and methods

2.1. *Image dataset*

The images in the database were captured using several different sensors (smartphones, compact cameras, DSLR cameras), and their resolutions range from 1 to 24 MPixels. Table 2 specifies the number of images for each plant/disease pair before (PDDB) and after (XDB) subdivision. Most disorders are related to fungi (77%), while 8% are due to viruses, 6% to various pests, 3% to bacteria, 2% to phytotoxicity, 2% to algae, 1% to nutritional deficiencies and 1% to senescence. Approximately 60% of the images were captured under controlled conditions, and 40% under real field conditions.

Some criteria were applied in order to make the image segmentation consistent. First, only images containing plant leaves were considered in the creation of the expanded dataset (XDB), because the symptoms appearing in other parts of the plants cannot always be suitably divided. The backgrounds of all images were manually blacked out prior to the subdivision. Five different types of signs and symptoms were identified, each one warranting different criteria for subdivision, always having diversity of characteristics as main goal. For all new images, healthy tissue occupied at least 20% of the total area in order to guarantee contrast with the diseased tissue. Because the criteria used to subdivide the images required human discretion to be applied correctly, the whole process was done manually.

The first type of symptom, scattered small, consists of numerous small lesions or spots spread over the leaf surface (Fig. 1a). Two criteria were used: relatively isolated symptoms were taken individually (Fig. 1b), and lesions that were part of clusters were taken as a group (Fig. 1c).

The second type of symptom, scattered large, consists of a number of large lesions or spots spread over the leaf surface (Fig. 2a). The criteria adopted here were the same as the previous case, but because of the larger symptom size, the box delimiting the area surrounding the lesion can potentially include parts of other lesions, even if they are relatively isolated. When this happened, in about half of the cases the

Table 2 – Image database composition with plant diseases and their hosts. The table does not include the 500 images of healthy tissue generated for each crop, as they were used only in part of the experiments.

Specimen	Disorder	Code	# Samples	
			PDDB	XDB
Common Bean	Anthracnose	A1	21	601
	Cercospora leaf spot	A2	2	41
	Cowpea mild mottle virus	A3	6	82
	Rust	A4	2	420
	Hedylepta indicata	A5	5	100
	Target leaf spot	A6	24	600
	Bacterial brown spot	A7	2	38
	Web blight	A8	7	75
	Powdery mildew	A9	12	183
	Phytotoxicity	A0	8	939
Cassava	Mites	B1	10	130
	Bacterial blight	B2	18	650
	White leaf spot	B3	9	115
Citrus	Algal spot	C1	5	249
	Citrus greasy spot	C2	10	271
	Canker	C3	9	227
	Citrus variegated chlorosis	C4	27	308
	Sooty mould	C5	4	148
	Leprosis	C6	18	65
	Halo blight	C7	5	7
	Mosaic of citrus	C8	15	440
	Scab	C9	2	153
Coconut tree	*Aspidiotus destructor*	D1	5	583
	Lixa grande	D2	33	609
	Lixa pequena	D3	34	195
	Cylindrocladium leaf spot	D4	5	100
	Phytotoxicity	D5	2	17
Corn	Anthracnose leaf blight	E1	3	12
	Anthracnose vein blight	E2	4	14
	Tropical corn rust	E3	14	889
	Southern corn rust	E4	15	3048
	Scab	E5	3	723
	Southern corn leaf blight	E6	44	3770
	Phaeosphaeria Leaf Spot	E7	31	779
	Diplodia leaf streak	E8	7	18
	Brown spot	E9	8	1071
	Northern corn leaf blight	E0	46	156
Kale	Alternaria leaf spot	F1	4	133
	Powdery mildew	F2	3	63
Cashew Tree	Algae	G1	3	111
	Anthracnose	G2	37	845
	Angular leaf spot	G3	8	1332
	Black mould	G4	33	2114
	Powdery mildew	G5	6	65
	Gummosis	G6	36	42
Coffee	Leaf miner	H1	12	52
	Cercospora leaf spot	H2	43	88
	Leaf rust	H3	17	383
	Bacterial blight	H4	37	1149
	Blister spot	H5	8	175
	Brown leaf spot	H6	25	52
Cotton	Seedling disease complex	I1	32	166
	Myrothecium leaf spot	I2	27	146
	Areolate mildew	I3	36	1711
Grapevines	Bacterial canker	J1	13	664
	Rust	J2	8	800
	Isariopsis leaf spot	J3	1	52
	Downy mildew	J4	22	597
	Powdery mildew	J5	29	134
	Fanleaf degeneration	J6	7	83

Table 2 – (*continued*)

Specimen	Disorder	Code	# Samples	
			PDDB	XDB
Passion fruit	Cercospora spot	K1	4	95
	Bacterial blight	K2	38	169
	Septoria spot	K3	7	16
Soybean	Bacterial blight	L1	56	3791
	Cercospora leaf blight	L2	5	10
	Rust	L3	65	2265
	Phytotoxicity	L4	23	1545
	Soybean Mosaic	L5	22	311
	Target spot	L6	62	966
	Downy mildew	L7	51	2306
	Powdery mildew	L8	77	1291
	Brown spot	L9	21	1248
Sugarcane	Orange rust	M1	18	1013
	Ring spot	M2	43	1656
	Red rot	M3	49	104
Wheat	Wheat blast	N1	14	82
	Leaf rust	N2	24	377
	Tan spot	N3	2	11
	Powdery mildew	N4	35	370
Total			1575	46,409

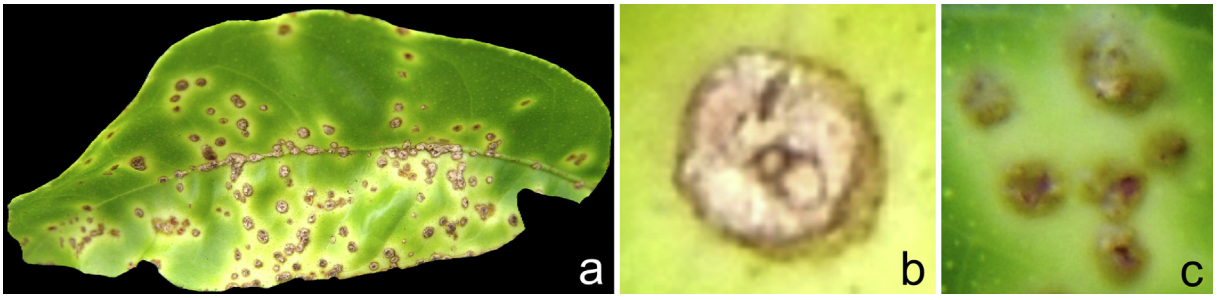

Fig. 1 – Example of scattered small symptoms (a), isolated lesion (b), and cluster of lesions (c).

Fig. 2 – Example of scattered large symptoms (a), spurious lesion blacked out (b), and spurious lesion unchanged (c).

spurious lesion was blacked out (Fig. 2b), and in the other half the spurious lesion was kept unchanged (Fig. 2c). This was done to increase the diversity of conditions.

The third type of symptom, isolated, consists of single lesions or spots (Fig. 3a). In this case, since there is only one region of interest, only one new figure is spawned from the original sample (Fig. 3b). The only exceptions to this rule were cases in which lesions were split into clearly distinct regions (Fig. 3c).

The fourth type of symptom, widespread, consists of large lesions that manifest over the entire leaf (Fig. 4a). Due to the wide variety of characteristics found in this group, the criteria for subdivision were only loosely defined. First, the entire original image (with the background removed) is also considered a sub-image. The remaining subdivisions were done by identifying relatively homogeneous regions within the diseased tissue (Fig. 4b,c).

Finally, the fifth type of symptom, powdery, consists of powdery spots on the leaf's surface (Fig. 5a). Those spots are

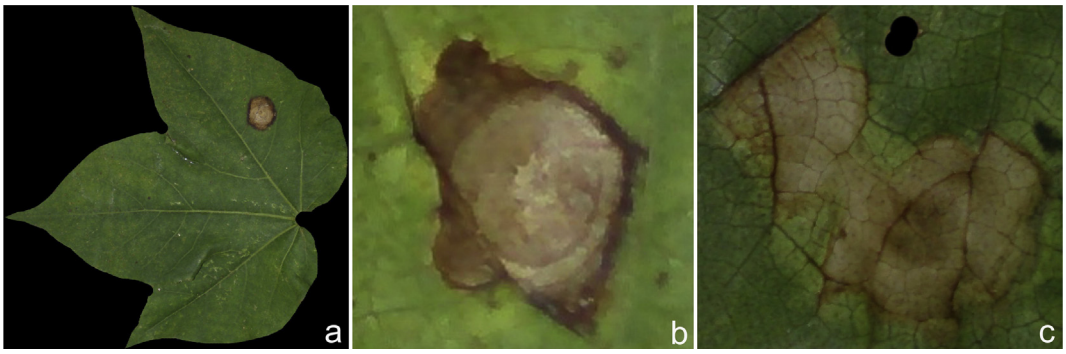

Fig. 3 – Example of isolated symptoms (a), lesion region delimited (b), and lesion with two visible regions (c).

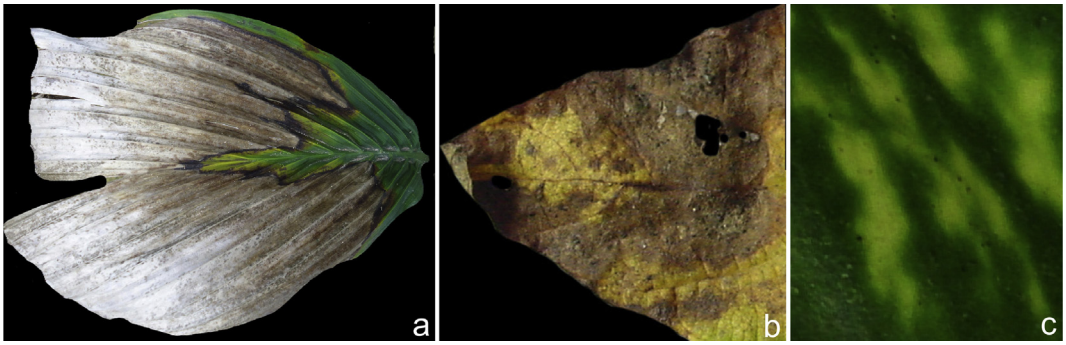

Fig. 4 – Example of widespread symptoms (a), lesions delimited by homogeneity (b and c).

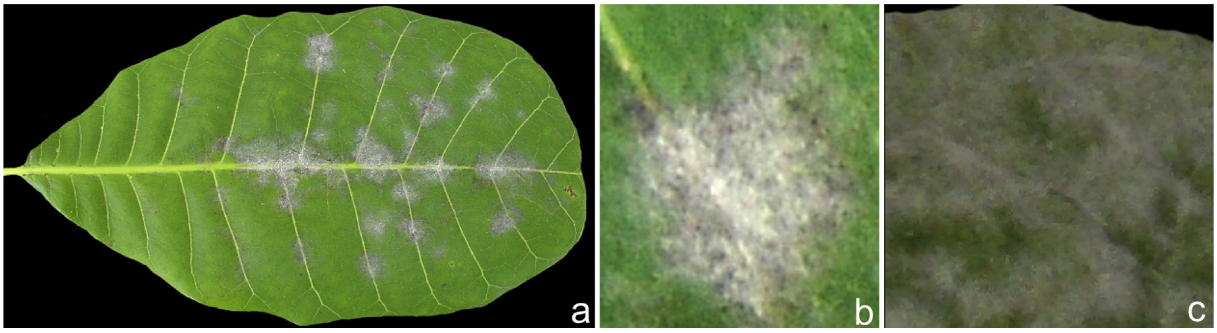

Fig. 5 – Example of powdery symptoms (a), isolated spots (b), and widespread spots (c).

initially separated, but as the disease evolves, they may merge into larger ones, until the leaf is completely covered. When spots were isolated, each one of them generated a new image (Fig. 5b). When the disease was more widespread, the division followed the same basic rules as widespread lesions (Fig. 5c).

2.2. *Experimental setup*

Transfer learning (Bengio, 2012) was applied to a pretrained GoogLeNet CNN using Matlab (Mathworks, Natick) Neural Network Toolbox (version 2017b). The GoogLeNet architecture was chosen because of its superior performance in the context of plant disease recognition (Ferentinos, 2018; Mohanty et al., 2016). The parameters used to train the network were the following: Base Learning Rate, 0.001; Momentum, 0.9; Mini Batch Size, 64; Number of Epochs, 5. These parameters were determined after a grid search in which different value combinations were tested on the soybean dataset. All experiments were run using a NVIDIA Titan XP Graphics Processing Unit (GPU).

Experiments were divided into two main groups. The first group focused on the classification problem, in which the objective was to determine the origin of a symptom that has already been observed and located. Thus, healthy samples should not be included in this set of experiments. For each crop, three sets of images were used: a) unmodified original images; b) original images with background removed; c) expanded dataset (XDB). In each case, 80% of the samples were used for training and 20% for validation. Additional CNNs were trained with a reduced version of the training dataset containing subdivided images, so it would match the size of the original training dataset. This was done as part of the investigation on whether training dataset size or using

localised regions was the main factor driving the improvement observed when XDB was used. Another experiment, aiming at evaluating the impact of severe class imbalance, was performed using only the expanded coffee dataset. In this experiment, the image set associated to each disease was augmented (Liu et al., 2018) until they all had the same number of samples (919) as the largest set (bacterial blight). This new set was used to train a CNN, and the results were compared with those obtained without any balancing attempt.

The second group of experiments focused on the detection problem, in which the objective was to detect disease signs amidst healthy tissue, for subsequent classification or not. Two experiments were carried out. The first one was similar to experiment c) in the first group, but with healthy samples included. The motivation was to determine if including the additional healthy samples would have significant impact on the classification accuracies. Actual detection was assessed in the second experiment, with the same CNN trained in the first experiment being used in this case. However, a new test dataset was generated, as follows. All images not used for training were rescaled to 2048 × 1368 pixels, in order to make the relative sizes of lesions roughly uniform across all images. A 224 × 224 sliding window was then applied to each image, with 50% overlap between consecutive frames, to generate a new set of cropped images. Images containing extraneous elements such as soil or background leaves were removed, and each remaining sample was labelled as healthy (no visible disease signs), mildly diseased (visible signs occupying less than 15% of the image), moderately diseased (visible signs occupying more than 15% and less than 50% of the image) and severely diseased (visible signs occupying more than 50% of the image). This labelling was done manually by visual inspection, so the division is not rigorous. However, since those labels were simply used to aggregate the images into groups roughly defined by the prominence of their symptoms, the subjectivity involved in the process was of little consequence. The new test dataset was then submitted to the trained CNN and two measures were calculated, the detection rate and the classification accuracy. Detection rate was given by the proportion of images not classified as healthy. The classification accuracy was given by the proportion of samples with symptoms correctly assigned, exactly as in all other experiments.

All experiments described in this section were carried out using a 10-fold cross-validation. Also, all images were resized prior to training to meet GoogLeNet's input dimension requirement (224 × 224 × 3 pixels). Although the network could be modified to accept other input sizes, the standard dimension was a good match to the typical image sizes in XDB. Augmentation techniques were applied to the training set in order to improve the robustness of the trained model (Liu et al., 2018).

3. Results

3.1. Classification experiments

Table 3 presents the overall accuracies obtained for each plant species, considering the original, background removed, and subdivided images with complete (C) and reduced (R) training datasets. Not all disorders shown in Table 1 were considered in the experiments with the original and background removed images, because some of them had too few images for a proper CNN training. With the exception of cotton and soybean, results were consistently improved by using images of individual lesions and spots. Accuracies for XDB tended to drop slightly when the reduced training dataset was used, indicating that the characteristics of the images were more relevant than the absolute number of samples used for training.

Because the number of classes and images and respective characteristics varied among crops, the effects of using the expanded dataset were also diverse. Figure 6 presents the confusion matrices associated to the expanded dataset for all crops, using the disorder codes presented in Table 1. An individual analysis for each crop is presented in the following.

Coffee: Results were consistent for all diseases, with exception of *Cercospora* leaf spot, which had an expressive

Table 3 – Accuracies obtained for each plant species (mean and standard deviation).

Crop	# Classes		# Images		Accuracy (%)			
	Original	Expanded	Original	Expanded	Original Images	Background Removed	Expanded	
							C	R
Common Bean	5	10	64	3079	83 ± 3.3	95 ± 0.8	94 ± 0.8	91 ± 1.6
Cassava	3	3	37	895	92 ± 2.8	83 ± 2.5	100 ± 0	100 ± 0
Citrus	7	9	87	1868	79 ± 5.9	62 ± 7.8	96 ± 0.6	93 ± 1.5
Coconut Tree	4	5	77	1504	97 ± 1.5	97 ± 0.4	98 ± 0.6	97 ± 1.2
Corn	7	10	165	10,480	60 ± 9.7	66 ± 8.0	75 ± 4.4	74 ± 6.5
Coffee	6	6	142	1899	76 ± 7.1	77 ± 5.2	89 ± 1.9	86 ± 2.5
Cotton	3	3	95	2023	100 ± 0	100 ± 0	99 ± 0.3	99 ± 0.5
Cashew Tree	3	6	78	4509	88 ± 3.2	83 ± 2.0	98 ± 0.5	96 ± 1.1
Grapevines	4	6	72	2330	75 ± 4.9	81 ± 2.7	96 ± 0.8	91 ± 3.0
Kale	0	2	0	196	–	–	100 ± 0	–
Passion Fruit	2	3	40	280	50 ± 12.6	90 ± 1.2	80 ± 4.2	80 ± 4.8
Soybean	8	9	377	13,733	82 ± 6.0	76 ± 6.3	87 ± 3.6	86 ± 4.1
Sugarcane	3	3	110	2773	93 ± 2.1	100 ± 0	99 ± 0.4	97 ± 1.0
Wheat	3	4	73	840	92 ± 2.9	61 ± 7.7	99 ± 0.5	98 ± 0.9
Total	*56*	*79*	*1383*	*46,135*	*82 ± 5.8*	*82 ± 4.5*	*94 ±2.0*	*91 ± 2.9*

Common Bean

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A0
A1	93	0	0	0	0	1	6	0	0	0
A2	0	88	0	0	0	0	0	12	0	0
A3	0	0	100	0	0	3	0	0	0	0
A4	0	0	0	100	0	0	0	0	0	0
A5	0	0	0	0	100	0	0	0	0	0
A6	0	0	0	0	5	85	0	0	10	0
A7	2	0	0	0	0	0	98	0	0	0
A8	0	5	0	13	0	0	0	100	0	0
A9	7	0	0	0	7	6	0	0	80	0
A0	0	0	0	0	0	0	0	0	0	100

Corn

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E0
E1	50	0	0	0	0	25	0	0	25	0
E2	0	100	0	0	0	0	0	0	0	0
E3	0	0	98	0	0	1	1	0	0	0
E4	0	0	0	97	1	2	0	0	0	0
E5	0	0	1	0	95	1	3	0	0	0
E6	0	0	0	2	1	96	1	0	0	0
E7	0	0	1	0	4	2	92	0	0	1
E8	0	0	0	0	0	75	0	25	0	0
E9	0	0	0	3	0	3	0	0	94	0
E0	0	0	5	0	0	35	5	0	0	55

Citrus

	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	100	0	0	0	0	0	0	0	0
C2	0	100	0	0	0	0	0	0	0
C3	0	0	98	0	0	0	0	2	0
C4	0	0	0	100	0	0	2	0	0
C5	0	0	15	0	77	0	0	8	0
C6	0	0	0	0	0	100	0	0	0
C7	0	2	2	0	0	0	96	0	0
C8	0	1	5	0	2	0	0	92	0
C9	0	3	0	0	0	0	0	0	97

Soybean

	L1	L2	L3	L4	L5	L6	L7	L8	L9
L1	94	0	1	2	0	1	0	0	2
L2	50	50	0	0	0	0	0	0	0
L3	0	0	98	0	0	2	0	0	0
L4	2	0	0	98	0	2	0	0	0
L5	0	0	1	0	95	2	2	0	0
L6	7	0	5	5	1	79	1	0	3
L7	0	0	0	0	0	0	100	0	0
L8	0	0	0	0	0	0	0	100	0
L9	14	0	0	5	0	12	0	0	69

Cashew Tree

	G1	G2	G3	G4	G5	G6
G1	96	4	0	0	0	0
G2	0	99	1	0	0	0
G3	0	0	97	3	0	0
G4	0	0	2	98	0	0
G5	0	0	0	0	100	0
G6	0	0	0	0	0	100

Coffee

	H1	H2	H3	H4	H5	H6
H1	82	0	0	18	0	0
H2	0	61	6	28	6	0
H3	0	1	99	0	0	0
H4	0	1	0	99	0	0
H5	0	0	9	0	91	0
H6	0	0	0	0	0	100

Grapevines

	J1	J2	J3	J4	J5	J6
J1	99	1	0	0	0	0
J2	0	100	0	0	0	0
J3	0	0	100	0	0	0
J4	0	0	0	99	0	1
J5	0	0	0	1	99	0
J6	0	0	0	3	0	97

Cassava

	B1	B2	B3
B1	100	0	0
B2	0	100	0
B3	0	0	100

Cotton

	I1	I2	I3
I1	97	0	3
I2	0	100	0
I3	0	0	100

Passion Fruit

	K1	K2	K3
K1	100	0	0
K2	12	88	0
K3	25	25	50

Coconut Tree

	D1	D2	D3	D4	D5
D1	100	0	0	0	0
D2	0	100	0	0	0
D3	0	0	100	0	0
D4	0	0	8	92	0
D5	3	0	0	0	100

Wheat

	N1	N2	N3	N4
N1	100	0	0	0
N2	4	96	0	0
N3	0	0	100	0
N4	0	1	0	99

Sugarcane

	M1	M2	M3
M1	100	0	0
M2	3	97	0
M3	0	0	100

Kale

	F1	F2
F1	100	0
F2	0	100

Fig. 6 – Confusion matrices obtained for all crops using the expanded dataset.

number of samples misclassified as Bacterial blight. The cause for those errors was the combination of similar symptoms with unbalanced representativeness (88 samples for the former, 1149 for the latter). In order to quantify the impact of this imbalance, an experiment in which all classes had the same number of training samples was performed (Section 2.2). Using balanced and imbalanced datasets resulted in almost the same accuracy (88.8% and 88.7%, respectively), but the errors caused by symptom similarity were much more equally distributed between the classes (Fig. 7).

Cassava: This crop had only a few samples associated, and the images of a given class tended to be captured in a single location, resulting in generally similar backgrounds. As a result, in many cases the CNN probably used the background,

	H1	H2	H3	H4	H5	H6
H1	82	0	0	18	0	0
	88	0	0	12	0	0
H2	0	61	6	28	6	0
	1	76	4	15	6	0
H3	0	1	99	0	0	0
	0	2	98	0	0	0
H4	0	1	0	99	0	0
	13	7	1	79	0	0
H5	0	0	9	0	91	0
	0	0	8	0	92	0
H6	0	0	0	0	0	100
	0	0	0	0	0	100

Fig. 7 – Confusion matrices obtained using the unbalanced (top values) and balanced (bottom values) coffee training dataset.

and not the symptoms, as the main distinctive feature. This is one of the potential negative effects of having an insufficient number of images for training and testing, and explains the seemingly poorer results when background was removed. It is worth noting that this crop had only three classes with relatively distinct symptoms, which made the task easier.

Common Bean: The accuracy for common bean using the expanded dataset was 11% higher than that obtained using the original images, but it was a little lower than that obtained using images with background removed. There are two explanations for this. First, many of the original images included whole diseased plants, rather than targeting a single leaf. Since leaves had different orientations and cast shadows on each other, it was difficult for the CNN to properly detect the features of interest in the images. When the background was removed from those images, only the centremost leaf was kept, making it easier to focus on the symptoms, greatly increasing the accuracy. Second, the number of classes considered for the expanded dataset was twice of that associated to the other two datasets, making the problem considerably harder. Thus, achieving accuracy of 94% with 10 classes can be considered an improvement. There were no diseases whose symptoms were consistently confused. The largest error rate, which was associated to powdery mildew (20%), was not caused only by symptom similarities with other diseases. Rather, because powdery mildew samples had similar lighting conditions to many other diseases, the network sometimes associated those specific conditions to different classes, especially when symptoms were slight or not very distinctive. This was one of the few cases for which augmentation operations were not able to prevent spurious model fitting.

Citrus: The results for citrus improved considerably after subdivision. The relatively poor results for the original set is due to the small number of images for testing, so even single misclassifications resulted in very large error rates. Results were even poorer for the background removed set, due to the tendency of the CNN to take the background into consideration when only a few images are available, as explained for cassava. The results for the subdivided datasets were consistently good, especially considering that nine different diseases were present. More importantly, most of the images were captured in the field under a variety of conditions. A major reason for this was that most diseases had hundreds of images associated, which translated to thousands of images available to be used in the training after dataset augmentation. The only diseases that had fewer images associated (Leprosis and Halo blight) have distinctive symptoms that could be learned by the network with just a few images.

Coconut tree: The results were also consistently good for this plant species. All five diseases had very distinctive symptoms among them, with the exception of Lixa Grande and Lixa Pequena, but the number of images was enough to avoid any major confusion.

Corn: As in the case of common bean, corn has 10 classes, but the results were significantly poorer, being the only crop for which the global accuracy was below 80%. This was expected, as the dataset associated to this crop includes many of the factors that may cause difficulties to any deep learning-based classifier (Barbedo, 2018a). The first and most obvious reason is the presence of diseases with similar symptoms. This is aggravated by the fact that the number of images for each disease is very imbalanced, varying from a dozen to a few thousands. This explains why most *Diplodia* leaf streak images were classified as Southern corn leaf blight – besides producing similar symptoms (Fig. 8), the number of images was more than 200 times larger for the latter. The second reason for the relatively large error rate was that almost all images for this crop were captured under uncontrolled conditions, often containing problematic illumination effects such as light-and-shadow and specular reflection (Fig. 9). Leaf regions severely affected by those phenomena were not removed, rather being included to stress the CNN capabilities. Experiments have shown that when the loss of information associated to those effects reached a certain amount, the network was no longer capable of reliably recognizing the symptoms.

Cotton: Although the accuracy for this crop was nearly perfect, the accuracy dropped slightly in comparison with the original dataset. The reason for this is that the shape of Seedling disease complex symptoms varies considerably, and in some specific cases they mimic the symptoms caused by Areolate mildew. If the entire leaf is considered, a variety of

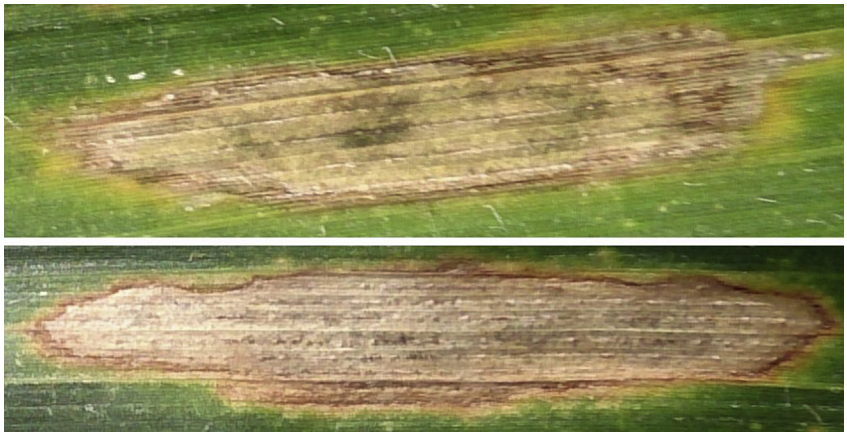

Fig. 8 – Example showing the similarity of symptoms produced by *Diplodia leaf streak* (top) and *Southern corn leaf blight* (bottom).

Fig. 9 – Image containing light-and-shadow and specular reflection effects.

shapes is considered together, making it easier to correctly recognise the disease.

Cashew tree: The results for this plant species were consistent for all datasets. The few errors that occurred were associated to the advanced stage of the diseases in some samples. In such cases, leaf tissue may begin to die, producing some characteristic visual cues that are only slightly related to the original symptoms of the disease that caused the necrosis.

Grapevines: Almost perfect accuracy was achieved for this crop, improving the original results in 21%. The few errors that occurred were mostly related to similarities between Downy mildew and Fanleaf degeneration.

Kale: Due to the small number of samples, this plant species was only considered in its expanded dataset version. Having to deal with only two classes, the CNN had no problem achieving 100% accuracy.

Passion Fruit: the difference in classification accuracies observed for the original (50%) and the background removed (90%) datasets was not entirely due to the busy backgrounds present in the former. As in the case of other crops, the small number of images used to validate the CNN caused individual errors to have a major negative impact on the accuracy percentages, which artificially decreased the accuracy obtained for the original dataset. In the case of the expanded database, many *Septoria* spot samples were classified as *Cercospora* spot and Bacterial blight, both due to symptom similarities and representativeness imbalance. No other major confusions were observed.

Soybean: Global results for soybean did not improve much (5%) when the expanded dataset was used. This was mainly due to the extreme imbalance between the number of samples available for *Cercospora* leaf blight (only 10) and for the other diseases (hundreds to thousands). This prevented the CNN to adequately learn the symptoms associated to this disease, leading to high error rates. All other errors were associated to diseases producing similar symptoms, such as Bacterial blight and Brown spot. The mild improvement using XDB was also associated to the fact that almost all images for this crop were captured under controlled conditions, making it easier for the CNN to learn patterns from complete images.

Sugarcane: The accuracies achieved for sugarcane were generally high. Most errors in the original dataset were due to busy backgrounds, while in the other datasets there were only a few sporadic errors. It is worth noting that almost all images associated to sugarcane were captured under controlled conditions, in which case CNNs tend to perform well with the original complete images.

Wheat: Results were significantly improved using XDB, with the few errors being mostly due to the heavily imbalanced number of samples between classes.

3.2. *Detection experiments*

For almost all crops, the inclusion of healthy samples had little impact on the model's effectiveness, with accuracies in most cases not deviating by more than 2% from the values presented in Table 3. Only three crops experienced higher drops in accuracy: cassava (100%–95%), kale (100%–89%) and wheat (99%–96%). On the other hand, the accuracy for corn rose from 75% to 78%. Most accuracy loss was caused by confusion between healthy samples and samples containing diseases with either powdery-like (Fig. 5) or mosaic-like symptoms. In general, considering that one additional class was added in all cases, the accuracy drop can be considered mild.

The effectiveness of the proposed approach in detecting diseased tissue is directly related to the prominence of the symptoms in the image (Table 4). While detection was most unsuccessful for mildly diseased images, success rates were much higher in the other cases, and the few moderately diseased samples that remained undetected had symptoms occupying less than 25% of the images. Therefore, if at least one quarter of the cropped image contains diseased tissue, successful detection is very likely. The proportion of false positives (healthy samples detected as diseased) was 11%. Most of those errors were due to the presence of dust, debris, water droplets and other extraneous elements. The classification accuracy obtained for samples detected as diseased was, in average, 10–12% lower than the accuracies obtained using the manually cropped images. This was expected, since the sliding window crops the images without any criterion regarding the position of the lesion in the frame. On the other hand, considering that each lesion will probably appear in multiple images, there will be many opportunities to get the classification right through a simple majority rule.

Table 4 – Detection rate and classification accuracy using the "sliding window" dataset. Two classification accuracies are presented: "all", which considers all samples, and "detected", which considers only samples correctly detected as diseased. The results obtained for all crops were aggregated for brevity.

Group	Detection Rate	Classification Accuracy	
		All	Detected
Healthy	–	89%	–
Mildly diseased	39%	31%	70%
Moderately diseased	97%	87%	80%
Severely diseased	100%	94%	85%

3.3. Processing times

Each training round using the original and subdivided sets of all crops took, in average, 13 min and 6.5 h, respectively. Considering the 10-fold cross-validation and the complete set of experiments, the total time spent training the CNNs was close to 140 h, using a single Nvidia Titan XP GPU. It is worth noting that training times can be greatly reduced by using multiple GPUs.

4. Discussion

CNN training may require a substantial number of images to yield reliable results, even with the application of transfer learning and augmentation techniques (Barbedo, 2018b; Kamilaris & Prenafeta-Boldú, 2018). This was the main motivation behind the creation of the expanded dataset. As explained in Section 2.1, the process of creating XDB was entirely manual, which took a few hundred hours to be completed. Fortunately, the approach of using localised images would not be labour intensive on the application side, as the widespread touchscreen technology enables the creation of applications allowing potential users to easily zoom in and select the region of interest. It is also worth noting that generating thousands of subdivided images requires fewer hours than capturing the same amount of new images. If automation of whole process is mandatory, using a sliding window to swipe the entire image can be a viable option. Results presented in Section 3.2 revealed that if symptoms fill at least 25% of the image, they will likely be detected, and since they may appear in multiple windows, correct classifications may be obtained by a majority rule. As a result, while the manual approach tends to be a little more accurate because it allows for symptoms to be carefully framed, the automatic approach can be advantageous under certain conditions.

The superior results obtained using the expanded dataset are directly associated to two main factors: 1) more images represent more opportunities for the neural network to learn the actual characteristics of the symptoms; 2) the elimination of spurious elements (e.g. background) by the segmentation process generated images with more homogeneous characteristics, allowing the network to focus on the right elements. The slight drop in accuracy observed when the number of training samples for XDB was reduced indicates that the second factor is more relevant. In fact, most additional errors were more related to diminished data variety than to the actual number of samples. The impact of reducing the training dataset was larger for crops with a wider variety of conditions.

One important consequence of having a larger and more homogeneous dataset is that the influence of images that are not good representations for the class will be weakened by the large number of proper samples, increasing the reliability of the training process. Thus, deleterious effects caused by the presence of spurious elements such as specular reflections and debris are greatly attenuated. This explains why using the expanded dataset was much more impactful when most of the images were captured under real conditions.

The ability of the expanded database to dilute deleterious effects caused by poor data is also useful in another context. Manual labelling of the images, being a subjective task, is prone to error. When only a few images are available for each class, wrong labels may have a substantial impact on the training process of the algorithms. Since XDB offers a large number of images for most diseases, the impact of wrong labels is considerably reduced.

Despite the superior performance achieved by using the expanded dataset, there are a few limitations that should be considered. After subdivision, the number of samples associated to each disease varied greatly due to the characteristics of the symptoms. Diseases that cause numerous small lesions or spots ended up having much more extensive sample collections. There are many factors that influence the ideal number of images that would be enough for the neural network to properly learn the characteristics of a disease's symptoms: intra-class symptom variability, diversity of conditions expected to be found in practice, similarity with other diseases, among others (Barbedo, 2018a). The experiments did not provide a clear answer on how many images would be enough for the neural network to properly learn the characteristics of its symptoms. In all cases, a few hundred images seemed to be enough to deliver reliable results, but this number has to be taken cautiously. Training and test datasets were taken from the same database, which contains only a very limited subset of all possibilities expected to be found in practice – as discussed in the Introduction, building a truly comprehensive database is currently unfeasible. Thus, it is possible to assert that a few hundred images are enough to properly deal with the conditions contemplated in the database used in the experiments, but it is not possible to claim that the trained CNNs will be robust under practical conditions. This may explain, at least in part, why Mohanty et al. (2016) observed a steep decrease in accuracy (from 99% to 31%) when their networks were applied to images that were not part of the original dataset.

Dataset representativeness has yet another layer that needs to be considered. The crops considered in this work had between 2 and 10 diseases associated. To the author's knowledge, there is no work in the literature that takes into consideration more than 10 diseases for a given plant species. The problem is that, in practice, each crop may have hundreds of disorders associated (Barbedo, 2016), from which only a very limited subset is usually considered. Even taking into account that such a subset usually contains the most common and economically important diseases, there will still be a vast amount of cases for which the CNN has not been trained, inevitably leading to misclassifications. This limitation is very difficult to overcome, because there are fewer opportunities to capture images from rarer disorders, and since they are lesser known, correctly labelling also becomes a challenge.

Many of the misclassifications observed when using the expanded dataset were due to severe imbalance regarding the number of samples used in the training. When diseases had relatively similar symptoms, and the number of samples used in the training was significantly different (more than 5-fold), the class with more samples was invariably favoured when there was some uncertainty associated to the visual characteristics displayed by the symptoms. To avoid this, it is

recommended that more samples be generated for the under-represented class or, if this is not possible, to reduce the number of training samples associated to other diseases.

Because transfer learning was used in association with a GoogLeNet CNN, all images used in this investigation were resized to 224 × 224 × 3 pixels. While the loss of information for the subdivided images was limited, as many of them already had small dimensions, such a reduction had significant impact on some of the original images. In some cases, it was observed that image resizing caused the resolution to fall below the level above which lesions could be properly resolved. This was the main factor behind the poor accuracies observed for wheat when complete images were considered. Thus, while the results using the expanded dataset were indeed superior, it is important to emphasise that part of the misclassifications associated to the complete images can be explained by the process of image resizing. Future research should employ more flexible network architectures capable of receiving images with different resolutions as input, thus avoiding the loss of essential information.

Although the experiments have indicated that image cropping can be an effective way to increase image datasets for plant images, it is important to emphasise that this procedure causes the loss of the information provided by patterns and density of spots on the leaves. This trade-off between data augmentation and loss of contextual information should be taken into consideration, because under certain conditions using the original images may be advantageous. Further investigations on this issue are expected to be carried out in the future.

This work was motivated by the lack of suitable datasets for proper application of machine learning techniques to plant pathology problems. The proposed approach, albeit imperfect, succeeds in mitigating the problem. However, other options to increase the amount of data available need to be further explored. An alternative that has already been successfully applied to a number of problems is the citizen science (Irwin, 2002). In this approach, non-professional volunteers collect and/or process data as part of a scientific enquiry (Silvertown, 2009). In the specific case of plant pathology, farmers and field workers could collect images in the field and, after uploaded to a server, those images would be properly labelled by an expert. This idea has been already carried out in practice, being the concept behind the commercial application Plantix™ (PEAT, Berlin). As mobile devices with imaging capabilities become ubiquitous, the challenge would be how to engage the farmers.

Another alternative to fill the data gap is by data sharing. There are research groups working on the automatic disease detection all over the world, using images collected in regions with very diverse characteristics. If the respective datasets were made available and properly integrated, the resulting set of images would be much more representative and research results would be more meaningful and applicable to real world conditions. As a step towards this goal, the dataset used in this work is being made available (https://www.digipathos-rep.cnptia.embrapa.br/). A further step would be to adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al., 2016). The generation of research data is usually expensive and time-consuming, and single use is thus a waste of resources that most often come from public funding. The FAIR principles imply that enabling and maximizing the fitness for reuse of research data involves much more than simply open access to data. The set of images could be published as a scientific citable product by itself, deposited in a long-term archive, described with detailed metadata and identifiable with a machine-readable digital identifier(s) that would be independent of the URL web-location. There are a few solutions that allow for those conditions to be met, such as Dataverse (Crosas, 2011) and ISA (Sansone et al., 2012).

5. Conclusion

The classification of plant diseases using digital images is very challenging. Deep learning techniques, and CNNs in particular, are seemingly capable of properly addressing most of the technical challenges associated to plant disease classification. On the other hand, dataset limitations in terms of both number and variety of samples still prevent the emergence of truly comprehensive systems for plant disease classification. Some efforts are underway towards building more representative databases, and data sharing is gradually becoming common practice, but the data available is still limited. The solution proposed in this article can not only increase the size of image datasets significantly, but can also increase the diversity of the data, as the natural variability within each image is indirectly taken into account by the subdivision into smaller regions. This approach also has some shortcomings, but it clearly leads to more reliable results in a context of limited data availability.

Acknowledgements

The author would like to thank Embrapa (SEG 02.14.09.001.00.00) for funding. The author would also like to thank Nvidia for donating the GPU used in the experiments.

REFERENCES

Amara, J., Bouaziz, B., & Algergawy, A. (2017). A deep learning-based approach for banana leaf diseases classification. In *Lecture notes in informatics (LNI)* (pp. 79–88). Bonn, Germany: Gesellschaft für Informatik.

Barbedo, J. G. A. (2013). Digital image processing techniques for detecting, quantifying and classifying plant diseases. *SpringerPlus, 2*, 660.

Barbedo, J. G. A. (2016). A review on the main challenges in automatic plant disease identification based on visible range images. *Biosystems Engineering, 144*, 52–60.

Barbedo, J. G. A. (2018a). Factors influencing the use of deep learning for plant disease recognition. *Biosystems Engineering, 172*, 84–91.

Barbedo, J. G. A. (2018b). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and Electronics in Agriculture, 153*, 46–53.

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. *Proceedings of the Workshop on Unsupervised and Transfer Learning, 27*, 17–37.

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

Crosas, M. (2011). The dataverse network®: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine, 17*(1).

Cruz, A., Luvisi, A., Bellis, L. D., & Ampatzidis, Y. (2017). X-FIDO: An effective application for detecting olive quick decline syndrome with deep learning and data fusion. *Frontiers in Plant Science, 8*, 1741.

DeChant, C., Wiesner-Hanks, T., Chen, S., Stewart, E. L., Yosinski, J., Gore, M. A., et al. (2017). Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning. *Phytopathology, 107*, 1426–1432.

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., & Park, D. S. (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors, 17*, 2022.

Hughes, D. P., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv, 1511*, 08060.

Irwin, A. (2002). *Citizen science: A study of people, expertise and sustainable development* (1st ed.). UK: Routledge, Abingdon-on-Thames.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70–90.

Liu, B., Zhang, Y., He, D., & Li, Y. (2018). Identification of apple leaf diseases based on deep convolutional neural networks. *Symmetry, 10*. Article 11.

Lu, Y., Yi, S., Zeng, N., Liu, Y., & Zhang, Y. (2017). Identification of rice diseases using deep convolutional neural networks. *Neurocomputing, 267*, 378–384.

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science, 7*. Article 1419.

Oppenheim, D., & Shani, G. (2017). Potato disease classification using convolution neural networks. *Advances in Animal Biosciences: Precision Agriculture, 8*, 244–249.

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., et al. (2012). Toward interoperable bioscience data. *Nature Genetics, 44*, 121–126.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology and Evolution, 24*(9), 467–471.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018.