Computers and Electronics in Agriculture 153 (2018) 46–53

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification

Jayme Garcia Arnal Barbedo

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

ARTICLE INFO

Keywords:
Image processing
Deep neural nets
Image database
Disease classification

ABSTRACT

The problem of automatic recognition of plant diseases has been historically based on conventional machine learning techniques such as Support Vector Machines, Multilayer Perceptron Neural Networks and Decision Trees. However, the prevailing approach has shifted to the application of deep learning concepts, with focus on Convolutional Neural Networks (CNNs). In general, this kind of technique requires large datasets containing a wide variety of conditions to work properly. This is an important limitation, given the many challenges involved in the construction of a suitable image database. In this context, this study investigates how the size and variety of the datasets impact the effectiveness of deep learning techniques applied to plant pathology. This investigation was based on an image database containing 12 plant species, each presenting very different characteristics in terms of number of samples, number of diseases and variety of conditions. Experimental results indicate that while the technical constraints linked to automatic plant disease classification have been largely overcome, the use of limited image datasets for training brings many undesirable consequences that still prevent the effective dissemination of this type of technology.

1. Introduction

The image-based classification of plant diseases is a difficult problem with a wide variety of challenges associated, including the presence of symptoms with extensive range of visual characteristics, possibility of multiple simultaneous disorders in a single plant, and different disorders having similar symptoms, among others (Barbedo, 2016). Extrinsic factors such as interference caused by the image background and illumination variations associated to capture conditions add even more complexity to the problem. While the combination of image processing and machine learning has led to many advances (Barbedo, 2013), practical use of tools like these has been limited. In the last few years, several studies have used the concepts of deep learning, and Convolutional Neural Networks (CNN) in particular, to try and make this kind of tool more accurate (Table 1).

Deep learning is a branch of machine learning composed by a number of algorithms that try to model high-level data abstractions using a deep graph with several processing layers containing linear and non-linear transformations (Goodfellow et al., 2016). Because CNNs have an intimate relationship between layers and spatial information, they are well-suited for image classification tasks (Arel et al., 2010), which explains their prevalence in recent plant disease classifiers. This type of neural network usually requires a very large number of samples for proper training, but this constraint can be relaxed by the application of transfer learning. This technique recycles previously trained networks by using the new data to update a small part of the original weights (Bengio, 2012).

Many of the studies found in the literature use transfer learning in their experiments (Mohanty et al., 2016; Brahimi et al., 2017; Ferentinos, 2018; Liu et al., 2018), and those that do not apply this technique use CNN architectures that are similar to existing ones (Amara et al., 2017; DeChant et al., 2017; Lu et al., 2017; Oppenheim and Shani, 2017). Also, many studies employed the initial PlantVillage dataset (Mohanty et al., 2016; Brahimi et al., 2017), which contains images that were mostly collected using a regularized process that generated relatively homogeneous backgrounds (Hughes and Salathé, 2015; Mohanty et al., 2016).

Thus, many studies are applying similar tools to a dataset that does not reproduce the range of conditions expected to be found in practice. This explains why most results reported in the literature show nearly perfect accuracy, without much variation between studies. It is quite revealing that when Mohanty et al. (2016) applied the model trained using the PlantVillage database to images originated from trusted online sources, the accuracy quickly fell below 50%. On the other hand, some studies applied their own datasets, but those were either collected under controlled conditions (Liu et al., 2018), and/or include only a few

E-mail address: jayme.barbedo@embrapa.br.

https://doi.org/10.1016/j.compag.2018.08.013
Received 30 March 2018; Received in revised form 10 July 2018; Accepted 5 August 2018
Available online 09 August 2018

Table 1
Studies employing deep learning for plant disease recognition. The accuracy is given by the number of samples correctly classified divided by the total number of samples.

Reference	CNN Network	Dataset	Accuracy	# Classes
Amara et al. (2017)	LeNet architecture	PlantVillage (extended)	92–99%	3
Brahimi et al. (2017)	AlexNet, GoogLeNet	PlantVillage	99%	9
Cruz et al. (2017)	Modified LeNet	Olive tree images (own)	99%	3
DeChant et al. (2017)	Pipeline	Corn images (own)	97%	2
Ferentinos (2018)	Several	PlantVillage (extended)	99%	58[a]
Fuentes et al. (2017)	Several	Tomato images (own)	83%	10
Liu et al. (2018)	AlexNet	Apple images (own)	98%	4
Lu et al. (2017)	AlexNet inspired	Rice images (own)	95%	10
Mohanty et al., 2016	AlexNet, GoogLeNet	PlantVillage	99%	38[b]
Oppenheim and Shani (2017)	VGG	Potato images (own)	96%	5

[a] Classes are distributed among 25 plant species.
[b] Classes are distributed among 14 plant species.

classes (Dechant et al., 2017; Fuentes et al., 2017; Lu et al., 2017). While all these studies yielded important contributions to the field, dataset limitations still prevent broader practical use.

This situation is in large part caused by the difficulties involved in building truly comprehensive databases. Most relevant visual manifestations of diseases happen in the field, as experiments with controlled inoculations often cannot produce the symptom variety found under more realistic conditions. Additionally, the visual characteristics of a symptom may change as the disease progresses and environmental factors such as humidity and temperature oscillate, so pictures may have to be taken frequently in order to cover the entire range of possibilities. It is also important to consider that all images need to be labeled with the correct disease, which is often a labor-intensive and error-prone process (Barbedo, 2018).

These circumstances require a better understanding about the effects of using relatively small datasets on the effectiveness of deep learning tools for plant disease classification. This is the objective and main contribution of this study. An image database, containing 12 plant species with very distinct characteristics in terms of number of samples and diseases, was used to test the behavior of CNN under a variety of conditions. The insights drawn from the experimental results led to a better understanding about the strengths and limitations of deep learning networks when these are trained with datasets of limited size and diversity. As a result, it was possible to draw some conclusions about the current development of deep learning-based plant disease classifiers, as well as to suggest some potential targets for future research on the subject. The database used in this work is being made freely available for academic purposes at a repository in the address https://www.digipathos-rep.cnptia.embrapa.br/.

2. Material and methods

2.1. Image dataset

The database available in the repository includes images of symptoms expressed not only on leaves, but also on stems, flowers and fruits. In this investigation, only images containing leaves were used in order to make the data more consistent. As a result, the image dataset used in this work is similar to the one used in Barbedo (2016). However, some diseases were removed from the original ensemble as they had too few images to be properly handled by CNNs, resulting in 56 diseases infecting 12 plant species. Since this dataset has already been detailed in Barbedo (2016), only a brief description is presented here. Additionally, only the common names of plants and diseases are presented; scientific names can be found in the database repository.

Table 2 shows how the database is distributed in terms of plant species and disorders. Images were captured using a variety of digital cameras and mobile devices, with resolutions ranging from 1 to 24 MPixels. About 15% of the images were captured under controlled conditions, and the remainder 85% of the images were captured under real conditions, with the leaves attached to the host plant. All images were stored in the 8-bit RGB format.

2.2. Experimental setup

Transfer learning (Bengio, 2012) was applied to a pretrained CNN (GoogLeNet) using the Neural Network Toolbox available in Matlab 2017b. The GoogLeNet architecture was chosen because of its superior performance in the context of plant disease recognition (Mohanty et al., 2016; Ferentinos, 2018). The parameters used to train the network were the following: Base Learning Rate, 0.001; Momentum, 0.9; Mini Batch Size, 16; Number of Epochs, 5. All experiments were run using a NVIDIA Quadro K620 Graphics Processing Unit (GPU).

In order to investigate the influence of the background on the results, two separate CNNs were retrained, the first using the original unprocessed images, and the second using whole images with background manually removed. In each case, 80% of the samples were used for training and 20% for validation. All images were resized prior to training to meet GoogLeNet's input dimension requirement ($224 \times 224 \times 3$ pixels).

In order to increase the size of the training set and decrease overfitting problems (Liu et al., 2018), the training datasets were augmented using a number of operations (Fig. 1).

The results are presented as confusion matrices with an overall accuracy associated (Table 3). The confusion matrices are given in terms of percentages, not absolute numbers. Those values were obtained using a 10-fold cross-validation. It is important to remark that the number of images, diseases and conditions for each plant species varies significantly. This allowed an investigation on the performance of the CNN under a wide range of different conditions and contexts.

3. Results

Table 3 presents the overall accuracies obtained for each plant species, considering the original and background removed images. Because background removal was explicitly investigated with separate CNNs, this factor was used to organize this section, following the four different behaviors that were observed: (a) no significant impact on the accuracies; (b) substantial accuracy improvement; (c) substantial accuracy decrease; (d) mixed results. Each subsection contains one pair of confusion matrices obtained for a selected plant species. The confusion matrices obtained for the other plant species are omitted due to space constraints.

3.1. Small background removal impact

Crops for which the impact of background removal was mild had in common the characteristic of having few classes (up to four) with

Table 2
Image database composition with plant diseases and their hosts.

Specimen	Disorder	# Samples
Common Bean	Anthracnose	21
	Powdery mildew	12
	Hedylepta indicata	5
	Target leaf spot	24
	Phytotoxicity	8
Cassava	Mites	10
	Bacterial blight	18
	White leaf spot	9
Citrus	Algal spot	5
	Chlorosis	27
	Canker	9
	Leprosis	18
	Bacterial spot	5
	Greasy spot	8
	Mosaic of citrus	15
Coconut tree	*Aspidiotus destructor*	5
	Lixa grande	33
	Lixa pequena	34
	Cylindrocladium leaf spot	5
Corn	Tropical corn rust	14
	Southern corn rust	15
	Southern corn leaf blight	44
	Phaeosphaeria Leaf Spot	31
	Diplodia leaf streak	7
	Brown spot	8
	Northern corn leaf blight	46
Cotton	Seedling disease complex	32
	Myrothecium leaf spot	27
	Areolate mildew	36
Coffee	Leaf miner	12
	Brown eye spot	43
	Leaf rust	17
	Bacterial blight	37
	Blister spot	8
	Brown leaf spot	25
Cashew Tree	Anthracnose	37
	Angular leaf spot	8
	Black mould	33
Grapevines	Bacterial canker	13
	Rust	8
	Downy mildew	22
	Powdery mildew	29
Soybean	Bacterial blight	56
	Rust	65
	Phytotoxicity	23
	Soybean mosaic	22
	Target spot	62
	Downy mildew	51
	Powdery mildew	77
	Brown spot	21
Sugarcane	Orange rust	18
	Ring spot	43
	Red rot	49
Wheat	Wheat blast	14
	Leaf rust	24
	Powdery mildew	35
Total		1383

relatively distinctive characteristics. In cases like this, it was observed that symptom differences outweighed deleterious effects of the backgrounds on the classification process (see Section 4). The confusion matrices obtained for coconut tree show an example of small background removal impact (Figs. 2 and 3). Specific remarks for each crop fitting this category are presented below.

Coconut tree has four classes with very unbalanced number of images. However, because the symptoms of the four diseases are mostly distinctive, error rates were low. The only exception was “Lixa grande” and “Lixa pequena”, which have relatively similar symptoms, but the small differences were correctly captured by the network. Background removal did not seem to have much impact, which may be explained by the fact that only “Aspidiotus destructor” had images captured in the field. This led to similar accuracies for the original and background removed images. In addition, training and test sets had relatively similar characteristics in terms of illumination and background, which helped the network to correctly capture symptom differences, instead of other spurious elements.

Cotton had only three classes, all producing symptoms with very distinctive characteristics. For that reason, the performance of the network was nearly perfect. A few errors occurred in the tests with background removed images. These were all concentrated in the “Myrothecium leaf spot”, which had some images in which the symptoms were still very mild and difficult to detect. Curiously, keeping the background in place seemed to help avoid those misclassifications. This means that the CNN is also using background information to perform the classification, which should not happen but it is an unavoidable consequence of having so few images to train and test the networks (Section 4). Again, no significant differences between training and test sets were observed.

The accuracies obtained for grapevines were identical for both sets of images, with only a few small differences in the distribution of errors. Almost all errors were due to confusion between powdery mildew and downy mildew.

As in the case of cotton, sugarcane had only three classes with very distinctive characteristics, resulting in very high accuracies. The few errors that occurred for the original images were associated to very busy backgrounds occupying a relatively large percentage of the images, and also due to a few test samples having illumination and background characteristics that were not represented in the training set.

3.2. Positive background removal impact

Background removal was expected to produce significantly higher accuracies in cases for which backgrounds tended to be busy and occupy a large portion of the images. This was indeed the case for a few crops, and the case of common bean is a very didactic example (Figs. 4 and 5). Most images for this crop were captured in the field and, due to the characteristics of the plants, the resulting backgrounds were, in general, very busy (Fig. 6). Because the symptoms produced by the diseases were reasonably distinctive, the accuracy improved greatly as soon as spurious elements were removed. Most of the variability between training and test samples was located in the backgrounds, which further explains the steep increase in accuracy.

The error rate for the cassava original images was a little lower because fewer classes were considered, but the deleterious effects of the background, together with a small number of images, was also very damaging. In the case of cashew tree, although there were only a few images with busy backgrounds, their removal resulted in nearly perfect accuracies. In both cassava and cashew tree cases, the few significant differences between training and test samples were related to background variability.

3.3. Negative background removal impact

In theory, removing the background should not deteriorate the classification performance. However, sometimes the CNN may actually use background characteristics to classify the images, instead of the symptoms themselves (Section 4). In this work, because the datasets were carefully built to avoid background overfitting, situations like this were not observed, except for one disease associated to soybean and another associated to corn (Section 3.4). However, background removal still had a negative impact for one plant species, but for different reasons.

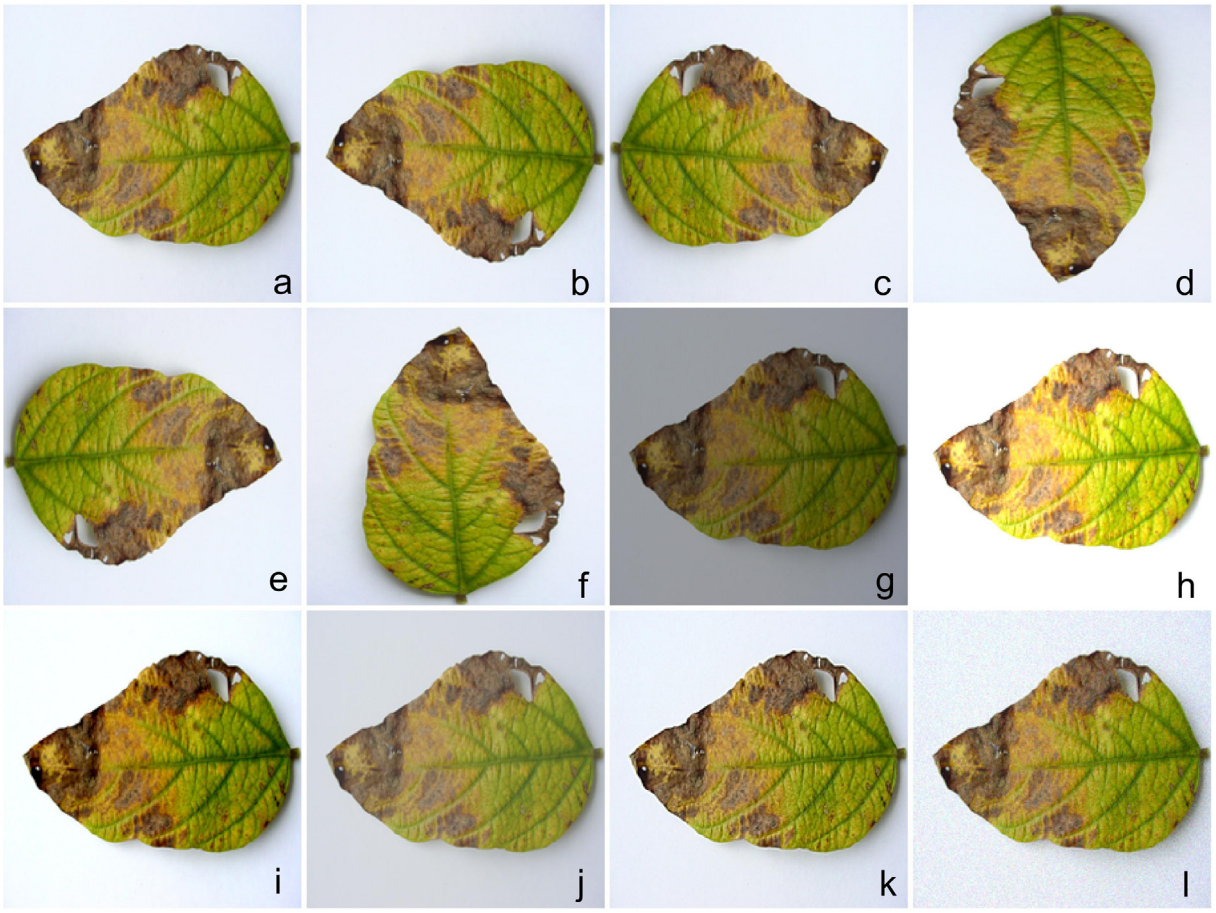

Fig. 1. Results of the augmentation process. (a) Original resized image; (b) vertical flipping; (c) horizontal flipping; (d) 90° counter-clockwise rotation; (e) 180° rotation; (f) 90° clockwise rotation; (g) random brightness decrease; (h) random brightness increase; (i) contrast enhancement; (j) contrast reduction; (k) sharpness enhancement; (l) addition of random Gaussian noise (0 mean, 0.001 variance).

Table 3
Accuracies obtained for each plant species. The accuracies are given by the number of samples correctly classified divided by the total number of samples.

Crop	Number of classes	Number of images	Accuracy	
			Original images	Back. removed
Common bean	5	64	65%	93%
Cassava	3	37	80%	100%
Citrus	7	87	88%	89%
Coconut tree	4	77	95%	96%
Corn	7	165	71%	73%
Coffee	6	142	82%	85%
Cotton	3	95	100%	98%
Cashew tree	3	78	93%	97%
Grapevines	4	72	90%	90%
Soybean	8	377	86%	78%
Sugarcane	3	110	97%	100%
Wheat	3	73	65%	50%
Total	56	1383	84%	87%

The classification performance for wheat was poor (Figs. 7 and 8). This was due to the close similarity between the symptoms produced by wheat blast and powdery mildew, which could not be resolved by the CNN. Because these two diseases were so similar, even when the CNN yielded a correct classification, the probabilities associated to both classes were always very close. When the background was removed, this seemed to have shifted the probabilities enough to significantly increase misclassifications. Thus, although in the wheat case the weight of the backgrounds was not high, even a small impact on the classification process was enough to alter the error rate considerably. There was also a high degree of confusion between rust and wheat blast, which was also caused by symptom similarities in the samples used in this work. It is worth noting that there are circumstances under which these three diseases can be more easily differentiated: differences between wheat blast and powdery mildew are very noticeable if other parts of the plant, like the spikes, are also taken into consideration, which was not considered in the context of this study; rust and wheat blast may produce dissimilar symptoms when the diseases are at different stages of development than those found in the database used in this study.

3.4. Mixed background removal impact

For some crops, removing the background resulted in improved accuracy for some classes, while at the same time increasing the error rates for others. This was observed for all crops having at least six diseases, as those have a wider variety of conditions and, as a consequence, end up experiencing the whole range of effects caused by background removal. Most observations made in Sections 3.2 and 3.3 hold in this case.

Citrus has a relatively high number of classes and a small number of images. The accuracies obtained for the original images and those with background removed was the same (Figs. 9 and 10), but sources of errors were different. Errors for the original images were mostly due to confusion between the classes “algal spot”, “canker” and “leprosis”. The symptoms produced by these diseases, although having some

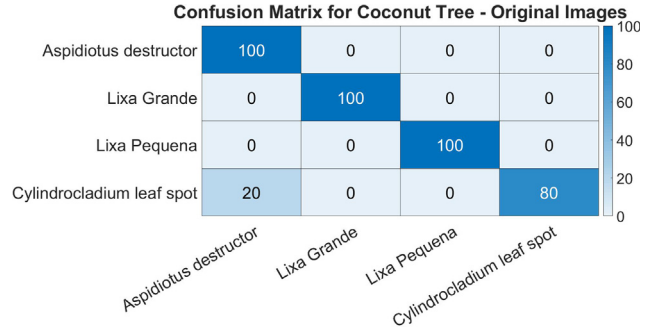

Fig. 2. Confusion matrix for coconut tree, using the original images.

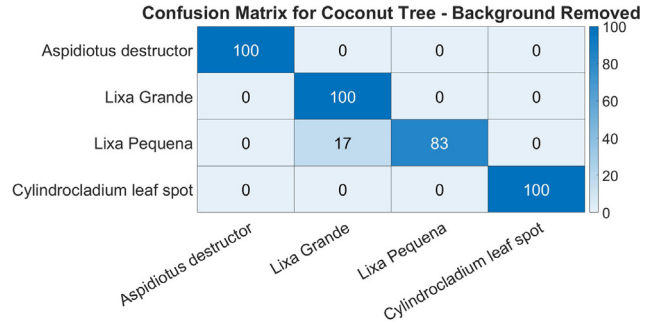

Fig. 3. Confusion matrix for coconut tree, using the images with background removed.

distinctive features, may have quite similar characteristics at certain stages of development. Curiously, although the symptoms remained the same, the confusion between “algal spot” and “leprosis” was eliminated when the background was removed, which is in large part explained by many test samples having backgrounds with features not found in the training dataset. On the other hand, there was a significant rise in the confusion between “leprosis” and “grease spot”. This happened because, in the original images, the backgrounds of the images used for validation were very different for these two classes, so it is likely that the CNN differentiated them using mostly the background information. Although this problem cannot be entirely avoided without complete background removal, a larger number of images with a wider variety of capture conditions could reduce its impact.

In the case of soybean, background removal had opposite effects in two classes. The accuracy for “phytotoxicity” improved from 75% to 90%, mostly because many of the images for this class had busy backgrounds (Section 3.2). On the other hand, the accuracy for “brown spot” fell from 77% to 4%. The reasons for this steep fall were exactly those discussed in Section 3.3. The “brown spot” class has many similarities with “bacterial blight”, but because all images of the latter were captured under controlled conditions, while the images of the former were more varied, background was an important discriminative factor in the original images. This highlights the impact that background may have over CNN results. The characteristics of soybean images in the training and test sets were relatively homogeneous, so there were not many errors caused by training and test mismatches.

Corn images were captured under a wider variety of conditions than any other crop. Since no selection was performed prior to the tests, many of the images included harmful effects such as light-and-shadow and specular reflections, which can greatly reduce the information available in the image (Barbedo, 2016). In addition, because certain conditions and characteristics were unique to single images, many of the samples in the test dataset had characteristics that the network could not handle properly, causing even more misclassifications. This resulted in lower overall accuracies, and also produced important differences between original and background removed images. As observed for soybean, two classes experienced opposite effects when background was removed. Tropical corn rust has characteristics that more or less match several other diseases, making this a difficult class to be correctly recognized. However, since all images in this class were captured under controlled conditions, all images whose background was more realistic would automatically be eliminated as a candidate to containing tropical corn rust symptoms, thus decreasing misclassifications. The “Brown spot” class had the opposite problem, as most images were captured in the field, producing very busy backgrounds. Thus, when the background was removed, error rates were naturally reduced.

Coffee also had two classes with opposite behaviors regarding background. Correct recognition of “Leaf miner” improved from 46% to 100% when background was removed. However, the backgrounds for this class were relatively homogeneous, so such an improvement was not expected. It was observed that this class often produced symptoms closely related to those associated to “brown eye spot”, which caused the CNN to assign close probabilities to these classes. Apparently, the act of removing the background “tipped the balance” in favor of the correct classification. The accuracy associated to “Leaf rust”, on the other hand, fell from 97% to 74% when the background was removed. This happened because, in many images, the symptoms produced by

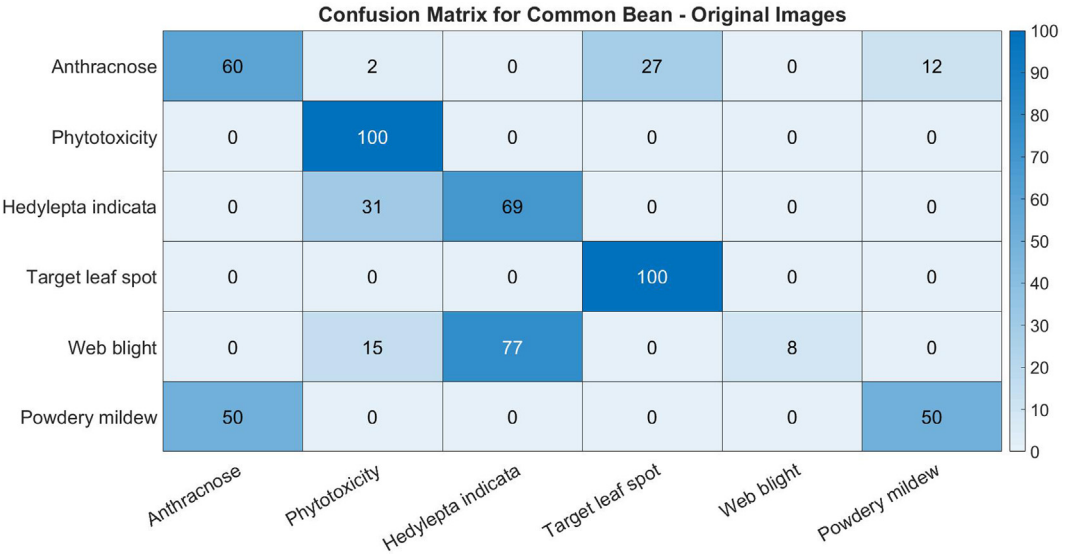

Fig. 4. Confusion matrix for common bean, using the original images.

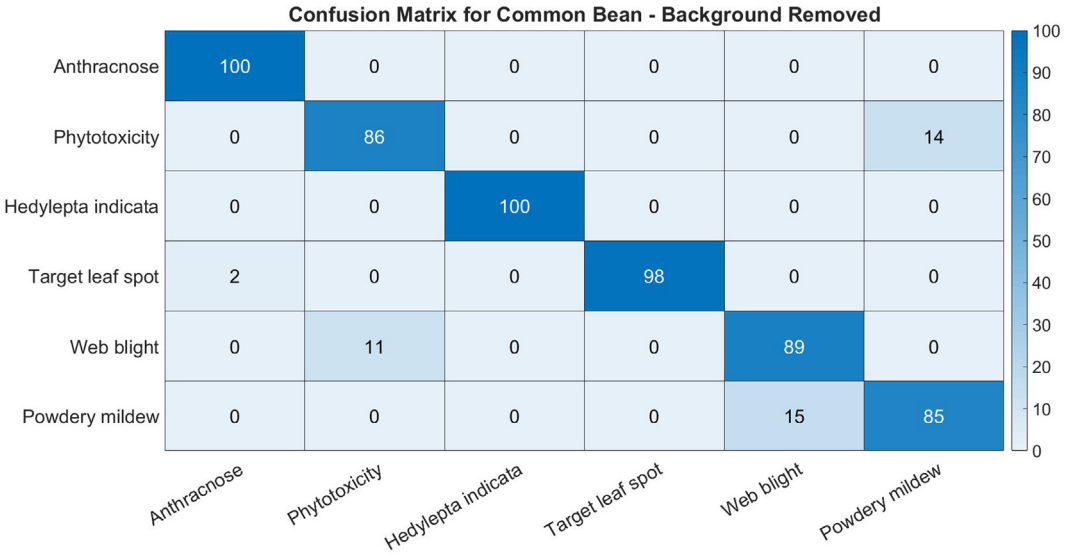

Fig. 5. Confusion matrix for common bean, using the images with background removed.

Fig. 6. Web blight symptoms (common bean) in an image with very busy background.

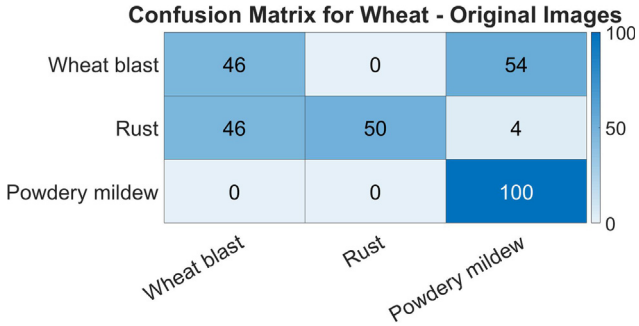

Fig. 7. Confusion matrix for wheat, using the original images.

rust were very mild, being difficult to detect even by human observers. In those cases, background played an important role on the identification, for the reasons discussed in Section 3.3.

4. Discussion

Even with the application of transfer learning and augmentation techniques, CNN training may require a substantial number of images to yield solid results (Kamilaris and Prenafeta-Boldú, 2018). However, defining how many images would be enough is not an easy task. The present investigation sheds some light onto the question, but a definite answer cannot be easily reached. The matter of fact is that each crop may have hundreds of disorders associated (Barbedo, 2016), so building a database that considers all of them is unpractical, not only because of the considerable effort required to capture the images, but also because those images would have to be correctly labeled, which is an even more difficult task (Kamilaris and Prenafeta-Boldú, 2018). Thus, the only way to deal with the problem is limiting the scope to a few more common diseases, having in mind that when an uncommon disease manifest, this will inevitably result in misclassification. It is also important to consider that more common diseases are in general more easily identifiable by farmers and farm workers, which means that a tool capable of identifying rarer disorders would be more useful.

Even with a limited scope, the number of images needed to take into consideration all possible capture conditions, symptom variations and sensor characteristics would still be impractically high. This means that any plant disease classifier will have associated a number of limitations that are directly related to the completeness (or lack thereof) of the dataset used to train it. A good example of the harmful effects of applying a network to images captured under different conditions was provided by Ferentinos (2018): when a CNN trained only with field-condition images was used to classify laboratory-condition images, the accuracy fell from 99% to 68%; when both types of images were reversed the accuracy fell to 33%.

Having incomplete databases is currently unavoidable, but many authors fail to include a discussion on this topic when presenting their findings. The dataset used in this study, while covering a wide range of image characteristics, has too few samples for the network to really capture the variety found in practice. This is reflected in the high error rates observed in some cases. Thus, while the networks trained in the context of this work are not ready for practical use, the results they yielded provided a wealth of information that can be explored in future developments. It is also worth noting that the database used here continues to receive new images, so its limitations tend to decrease with time.

A few useful remarks can be drawn within the scope delimited by the dataset used here. The accuracies presented in Section 3 were obtained by feeding the trained CNN with images that were not used for training. The differences observed between plant species were basically due to three main factors: (1) Number of classes considered; (2) Similarity of characteristics between the images present in the training and test datasets; (3) characteristics and variability of the image backgrounds. The association of accuracy with the number of classes is clear: as this number grows the chances of choosing the wrong class rises due to the increased number of options and to the higher

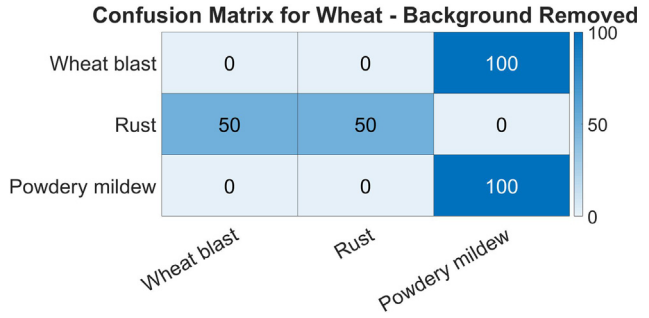

Fig. 8. Confusion matrix for wheat, using the images with background removed.

probability of existing pairs of diseases with similar symptoms.

CNN networks only learn the characteristics and features present in the training dataset. If the characteristics of the images used for testing depart too much from those found in the training set, accuracy is expected to be low. When the training set is truly comprehensive, this problem is diminished as the network is prepared to deal with a wide variety of images, so high accuracies can usually be expected. However, considering that most training datasets have important limitations, the accuracies obtained for a given set of images will be highly dependent on how their characteristics compare with those learnt by the CNN. One important consequence of this observation is that even if the training and test datasets are carefully chosen, none of them will include all possible data variety, which in turn means that the obtained accuracies will reflect, in large part, the similarity (or lack thereof) between the characteristics found in both datasets. Hence, all results reported in the literature are only valid for the particular test sets used in those experiments. When the network classify new samples, accuracies may either increase, if the characteristics of the new images are closer to those found in the training set, or decrease, if those characteristics are more diverse. The latter is usually more likely because, with a few exceptions (Cruz et al., 2017), both training and test datasets come from the same database, likely resulting in a relatively high degree of similarity between samples in both sets. This may explain, at least in part, why Mohanty et al. (2016) observed such a steep decrease in accuracy (from 99% to 31%) when their networks were applied to images that were not part of the original dataset. Thus, it is recommended that all results reported in the literature be interpreted in light of the representativeness limitations associated to the respective training datasets.

Ideally, image backgrounds should not contribute for the classification, as these do not change as different diseases are considered. However, it is possible to draw from the results reported in Section 3 that the background may have some serious impact on the results produced by a CNN, especially in the case of small datasets whose images were captured under realistic conditions. This is in stark contrast with the findings reported by Mohanty et al. (2016), who observed that leaf segmentation actually degraded the results. However, these authors used images collected using a regularized process that generated relatively homogeneous backgrounds. Other investigations reported in the literature often ignore this issue, either because the images were captured under controlled conditions, or because the authors did not deem this issue important. This may lead to unrealistic results that do not apply in practice. If the database is large enough to include a wide variety of backgrounds for all classes being considered, the influence of spurious objects may become diluted, but given the difficulties involved in building a truly comprehensive plant disease image database, this is generally not the case.

5. Conclusion

This paper presented an investigation on the application of the concepts of deep learning and transfer learning to the problem of plant disease classification. The study was carried out using a dataset that, while varied in terms of plant species, diseases and image capture conditions, had a number of samples that often was too small for the CNN to thoroughly capture the characteristics and variations associated to each class. While this prevents the trained networks to be used in practice, those limitations produced a wealth of information that can be used in future studies on the subject.

The main conclusion that can be drawn from this study is that CNNs are indeed powerful tools that can suitably deal with plant pathology problems. The main limitation that still prevents this kind of tool to be more widely used in practice is not technical, but practical. Building databases comprehensive enough for the creation of truly robust tools is very challenging. Some initiatives are using the concepts of social networks to accelerate the process (Barbedo, 2018), but there is still much to be done. The good news is that many groups are pursuing this goal, and collaborations and data sharing are becoming common practice. In this context, the database used in this work is being made available, in the hopes that this will help to further advance the research on this subject.

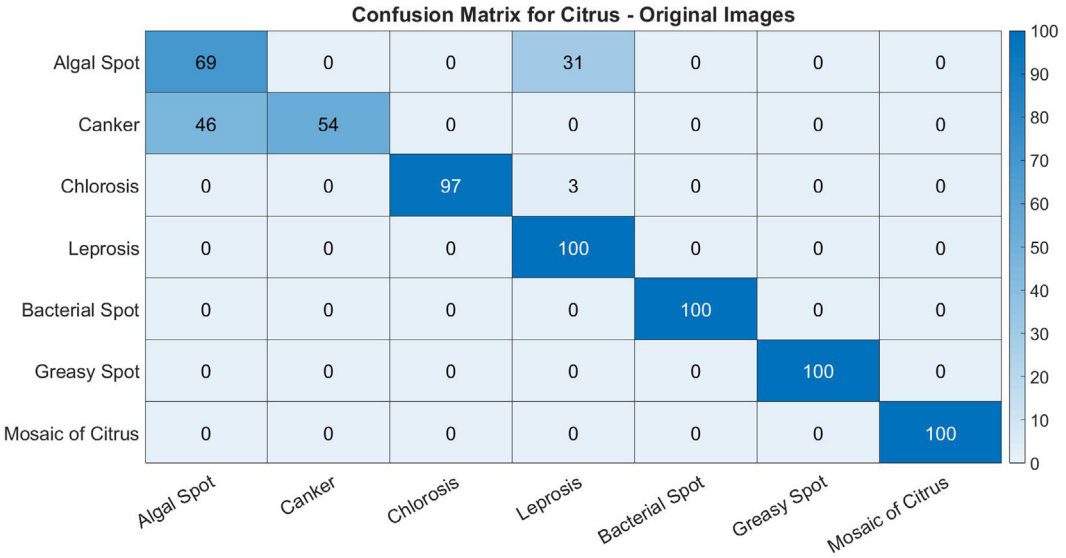

Fig. 9. Confusion matrix for citrus, using the original images.

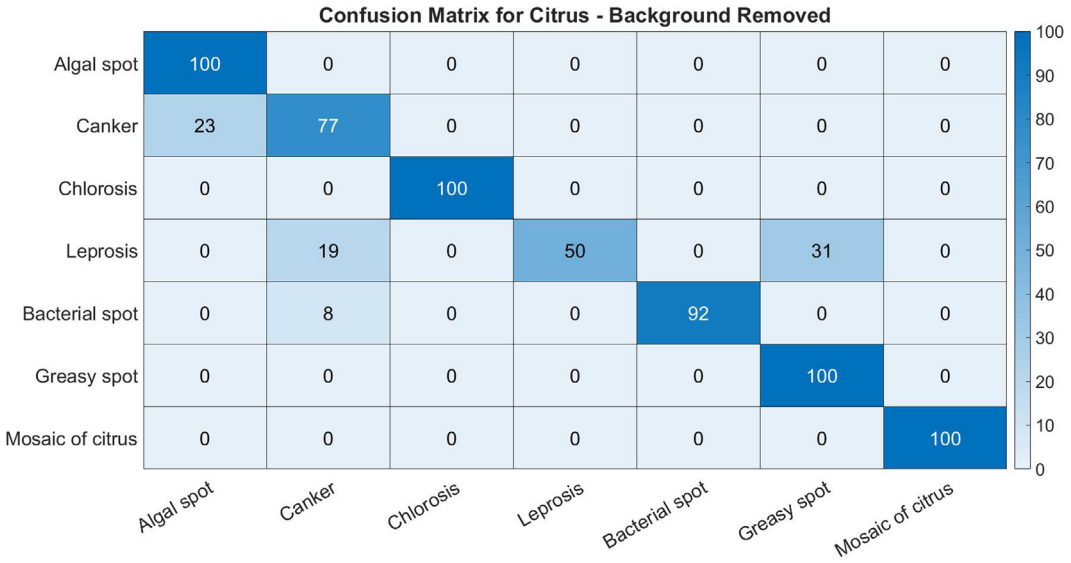

Fig. 10. Confusion matrix for citrus, using the images with background removed.

Acknowledgements

The authors would like to thank Embrapa (SEG 02.14.09.001.00.00) for funding.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compag.2018.08.013.

References

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

Arel, I., Rose, D.C., Karnowski, T.P., 2010. Deep machine learning - a new frontier in artificial intelligence research. IEEE Comput. Intell. Mag. 5 (4), 13–18.

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. Biosyst. Eng. 144, 52–60.

Barbedo, J.G.A., 2018. Factors influencing the use of deep learning for plant disease recognition. Biosyst. Eng. 172, 84–91.

Bengio, Y., 2012. Deep Learning of Representations for Unsupervised and Transfer Learning. Proc. Workshop Unsupervised Transf. Learn. 27, 17–37.

Brahimi, M., Boukhalfa, K., Moussaoui, A., 2017. Deep learning for tomato diseases: classification and symptoms visualization. Appl. Artif. Intell. 31 (4), 299–315.

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. Front. Plant Sci. 8, 1741.

DeChant, C., Wiesner-Hanks, T., Chen, S., Stewart, E.L., Yosinski, J., Gore, M.A., Nelson, R.J., Lipson, H., 2017. Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning. Phytopathology 107 (11), 1426–1432.

Ferentinos, K.P., 2018. Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. 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.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, MA.

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.

Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: a survey. Comput. Electron. Agric. 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, 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. Front. Plant Sci. 7, 1419.

Oppenheim, D., Shani, G., 2017. Potato disease classification using convolution neural networks. Adv. Anim. Biosci.: Prec. Agric. 8 (2), 244–249.