biosystems engineering 172 (2018) 84–91

Available online at www.sciencedirect.com

ScienceDirect

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

Research Reassessment

Factors influencing the use of deep learning for plant disease recognition

Jayme G.A. Barbedo

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

ARTICLE INFO

Article history:
Received 22 March 2018
Received in revised form
25 April 2018
Accepted 30 May 2018
Published online 17 June 2018

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

Deep learning is quickly becoming one of the most important tools for image classification. This technology is now beginning to be applied to the tasks of plant disease classification and recognition. The positive results that are being obtained using this approach hide some issues that are seldom taken into account in the respective experiments. This article presents an investigation into the main factors that affect the design and effectiveness of deep neural nets applied to plant pathology. An in-depth analysis of the subject, in which advantages and shortcomings are highlighted, should lead to more realistic conclusions on the subject. The arguments used throughout the text are built upon both studies found in the literature and experiments carried out using an image database carefully built to reflect and reproduce many of the conditions expected to be found in practice. This database, which contains almost 50,000 images, is being made freely available for academic purposes.

1. Introduction

Since the dawn of agriculture, plant diseases cause substantial economic, social and environmental losses. Prophylactic actions are not always enough to prevent apidemics, thus careful monitoring is essential for early detection and consequent application of control measures. Traditionally, crop inspection has been carried out visually by people with some training or experience detecting plant disorders. As for any activity carried out by humans, this approach is subject to psychological and cognitive phenomena that may lead to bias, optical illusions and, ultimately, to error (Bock, Poole, Parker, & Gottwald, 2010). More importantly, trained plant pathologists are not always available, especially in poor and isolated areas. It is also worth noting that many agricultural areas are too expansive to be properly monitored throughout (Barbedo, 2013). Image-based tools can thus play an important role in detecting and recognising plant diseases when human assessment is unsuitable, unreliable or unavailable.

At present, the potential of automated tools has yet to be realised. In previous works, we have investigated the characteristics of the proposals found in the literature (Barbedo, 2013) and the main challenges that still prevent this kind of technology from being adopted in practice (Barbedo, 2016). Methods based on conventional machine learning techniques have been relatively successful under limited and constrained setups, but many of the difficulties associated with the intrinsic characteristics of the problem could not be properly handled. With the inception of deep learning concepts, the answer to those limitations seemed to be close. Indeed, since 2015 research on plant disease detection has strongly veered towards using deep learning.

According to Ferentinos (2018), deep learning refers to "the use of artificial neural network architectures that contain a quite large number of processing layers, as opposed to

E-mail address: jayme.barbedo@embrapa.br.
https://doi.org/10.1016/j.biosystemseng.2018.05.013
1537-5110/

shallower architectures of more traditional neural network methodologies". Among deep learning tools, arguably the most commonly used are the Convolutional Neural Networks (CNN) (Krizhevsky, Sutskever, & Hinton, 2012). This kind of neural network requires fewer artificial neurons than conventional feedforward neural networks, being particularly suitable for image recognition. CNNs usually require a very large number of samples to be trained; however, in many real-world applications, it is expensive or unfeasible to collect the training data needed by the models (Pan & Yang, 2010). Thus, many authors are applying the concept of transfer learning to reuse pretrained networks (e.g. GoogLeNet and AlexNet), in which case predictions are done on examples that are not from the same distribution as the training data (Bengio, 2012). The conjunction of deep learning and transfer learning, together with the development of Graphics Processing Units (GPU), has provided a powerful tool for classification and recognition of diseases in plants (Ferentinos, 2018).

The application of deep learning to plant pathology problems started to gain momentum after 2015 (Table 1). As promising as the results seem to be, they must be interpreted with some observations in mind: a) most of the studies cited above use transfer learning in their experiments (Brahimi, Boukhalfa, & Moussaoui, 2017; Cruz, Luvisi, Bellis, & Ampatzidis, 2017; Ferentinos, 2018; Liu, Zhang, He, & Li, 2018; Mohanty, Hughes, & Salathé, 2016), and even those that do not apply this technique use CNN architectures that are similar to existing ones (Amara, Bouaziz, & Algergawy, 2017; DeChant et al., 2017; Lu, Yi, Zeng, Liu, & Zhang, 2017; Oppenheim & Shani, 2017); b) many studies used images contained in several versions of the PlantVillage dataset (Amara et al., 2017; Brahimi et al., 2017; Cruz et al., 2017; Ferentinos, 2018; Mohanty et al., 2016). As a consequence, most studies are applying similar tools to similar datasets. So, it is no surprise that there is not much variation in the results reported in the literature. In addition, there are many factors that affect deep learning-based tools when they are used under real field conditions, but in most cases these factors are only briefly discussed (or not considered at all). This causes the practical use of tools for automatic disease recognition to be still very limited. Although initiatives such as Plantix (PEAT, Berlin, Germany) are trying to change this scenario, there is still much to be done in order to effectively introduce this kind of technology to the daily routine of farms.

This article provides an in-depth analysis of the main factors that affect the performance of deep learning-based tools for plant disease recognition under realistic conditions. The goal was to provide some guidelines to make the investigation of deep learning-based methods for disease recognition more thorough and realistic.

The relevant factors mentioned above and discussed in detail in Section 4 were derived from experiments using CNNs. This was done by carefully analysing each misclassification produced by the model, and then associating them with a specific causal factor. This analysis provided a wealth of information which was used to draw out the remarks associated with each one of those factors. When appropriate, causes of misclassification were linked to error sources reported in the literature, thus providing further evidence of their generality. Thus, in the context of this work, the absolute accuracies yielded by the model are not nearly as important as the underlying causes for the errors. It is also important to emphasise that the image database used in the experiments was carefully built to reflect and reproduce many of the conditions expected to be found in practice. This database, which contains almost 50,000 images, is being made freely available for academic purposes at the address https://www.digipathos-rep.cnptia.embrapa.br/.

2. Materials and methods

The database used in the experiments is freely available and contains almost 50,000 images of 171 diseases affecting 21 plant species. However, only images of corn diseases were used in the context of this work. This subset was chosen because it contains the widest variety of conditions and a reasonable number of images for each of the nine diseases (Table 2), all of which are caused by fungi.

One of the main advantages of the deep learning approach is that, in general, the symptoms do not have to be explicitly identified in the image. However, relevant information is mostly concentrated in the symptoms themselves and their surrounds. Thus, in order to increase the size of the database and to test how the CNN would perform with more localised information, the original samples were divided into smaller images containing individual lesions or localised symptom regions (Fig. 1).

Some rules were applied for consistency in this division: a) images were manually blacked out prior to the subdivision; b) healthy tissue occupied at least 20% of the cropping area; c) isolated symptoms were taken individually; d) clustered

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 et al. (2018)	CNN (AlexNet)	Apple images (own)	98%
Lu et al. (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%

Table 2 – Corn leaf samples used to train and test the CNN used in this work.

Disease (common name)	Disease (scientific name)	Number of samples (original)	Number of samples (after subdivision)
Anthracnose	*Colletotrichum graminicola*	7	26
Tropical rust	*Physopella zeae*	14	889
Southern corn rust	*Puccinia polysora*	15	3051
Scab	*Gibberella zeae*	3	723
Southern corn leaf blight	*Bipolaris maydis*	44	3773
Phaeosphaeria Leaf Spot	*Phaeosphaeria maydis*	31	779
Diplodia leaf streak	*Stenocarpella maydis*	7	18
Physoderma brown spot	*Physoderma maydis*	8	1072
Northern Leaf Blight	*Exserohilum turcicum*	46	110

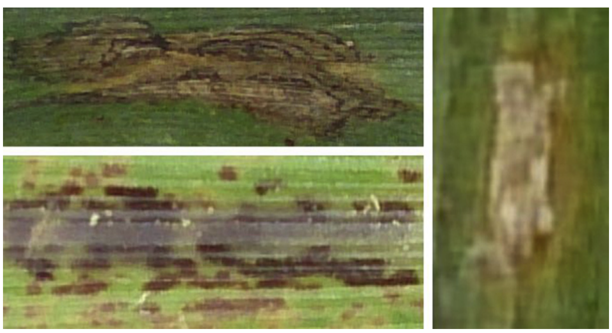

Fig. 1 – Examples of subdivided images. Top left: anthracnose; bottom left: *physoderma* brown spot; right: tropical rust.

symptoms were taken as a group; e) widespread symptoms were taken both as a whole and divided into homogeneous regions.

Transfer learning was applied to a pretrained CNN (GoogLeNet) using the Neural Network Toolbox provided by Matlab 2017b. At first, three different CNNs were trained, the first using the original unprocessed images, the second using whole images with background manually removed, and the third using the subdivided images. In each case, 80% of the samples were used for training and 20% for validation. A fourth CNN was trained with a reduced version of the training dataset containing subdivided images, so it would match the size of the original training dataset. This last CNN was used to investigate the relative impact of different factors associated with the expanded dataset (Section 4.6). Training datasets were augmented using operations of rotation, mirroring, addition of Gaussian noise, brightness adjustment and contrast adjustment. As a result, the size of the training set was increased 12-fold and overfitting problems were reduced (Liu et al., 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).

The final accuracies for each network were obtained using a 10-fold cross-validation approach. The characteristics of the images that were misclassified by the CNNs were carefully analysed, and nine factors were identified as having the most impact on the results. Those factors are introduced in Section 3 and discussed in detail in Section 4.

3. Results

As mentioned in the introduction, absolute accuracies are secondary in the context of this work. However, relative differences between the four trained neural networks are important indicators of some of the main factors that affect the effectiveness of CNNs for plant disease recognition (Sections 4.4 and 4.6). Table 3 shows the accuracies obtained using each trained CNNs. The number of training samples includes the images generated by augmentation techniques.

Careful analysis of the circumstances under which misclassifications tended to occur revealed that there are nine major factors that affect the performance of CNNs for plant disease recognition. Those factors are listed and briefly described in Table 4, and discussed in detail in Section 4.

4. Discussion

Barbedo (2016) discussed a number of intrinsic and extrinsic factors that affect the automatic recognition of plant diseases. Such a discussion was built at a time when the application of deep learning to the problem was still nascent. Those factors are now revisited under the perspective of deep neural networks and their peculiarities. New aspects, derived both from the literature and from the practical experiments, were also included. Extrinsic factors are discussed in Sections 4.1–4.5, and intrinsic factors are discussed in Sections 4.6–4.9. Section 4.10 presents some additional minor factors.

4.1. *Annotated datasets of insufficient size and variety*

Many of the problems discussed in this article are at least partially due to limitations in the dataset used to train the CNNs. This is a fact recognised by most authors, but not always properly taken into account in the analysis of the results.

Table 3 – Accuracies obtained using CNNs trained with different datasets.

Dataset	# Training samples	Accuracy
Original	1584	76%
Background removed	1584	79%
Subdivided (full)	100,608	87%
Subdivided (reduced)	1584	81%

Table 4 – Factors impacting the performance of CNNs for plant disease recognition. The first five factors are extrinsic to the disease recognition problem, and the last four are intrinsically connected to the issue.

Factor	Impact
Limited annotated datasets	Datasets do not have enough samples for deep neural networks to properly learn the classes. Annotation errors may damage the learning process.
Symptom representation	Datasets do not adequately represent the symptom variety found in practice, weakening the robustness of the trained model.
Covariate shift	Training and testing a model using the same dataset often leads to unrealistic performance assessment, as the model will likely fail when applied to other datasets.
Image background	Image background may contain elements that may disturb the training process, especially if those elements are present in multiple samples.
Image capture conditions	Images can be captured in a wide variety of conditions. In order to be representative, a dataset has to contemplate all possibilities, which is currently unfeasible.
Symptom segmentation	Images with many spurious elements often cause difficulties for the model. Taking more localised regions of the leaf may prevent some problems.
Symptom variations	Symptoms produced by a disease may present a wide range of characteristics. It is difficult to build datasets capable of representing symptom diversity properly.
Simultaneous disorders	It is difficult to detect multiple simultaneous disorders when images are analysed as a whole. Adopting localised symptom regions may mitigate this problem.
Disorders with similar symptoms	Some disorders produce visually similar symptoms. In cases like this, simple RGB images may not be enough for proper recognition, even with well-trained models.

The use of transfer learning, in which only a few layers of pre-trained neural networks are adjusted to new datasets, has greatly reduced the need for massive datasets. Also, data augmentation techniques can artificially expand some training datasets using label-preserving transformations. Nevertheless, at least a few hundred images are always required (Kamilaris & Prenafeta-Boldú, 2018) and, depending on the complexity of the problem, the number may be much higher.

Capturing the images is only part of the problem. Data annotation is a necessary operation in the large majority of cases (Kamilaris & Prenafeta-Boldú, 2018). In many situations, this task has to be carried out by experts. This is particularly true in the case of plant pathology, as it is not always trivial for a non-expert to determine which disease is present in a plant. More importantly, in a few cases not even an expert is capable of providing a definite label without laboratory tests. This has two consequences: 1) rigorous labelling of images is a slow and expensive process; 2) labels may have some degree of uncertainty associated. If ground-truth data is not reliable, the training process will not be adequate. This is concerning, especially considering that some initiatives are using the concepts of social networks to build databases large enough to allow the development of truly robust tools (Barbedo, 2017). Since in those arrangements the labelling is usually based only on visual cues present in the images, mislabelling may occur more often. Unfortunately, building extensive plant disease databases with rigorous labelling is currently impractical. However, if the number of mislabelled samples is not too high, there is some evidence that deep neural networks are robust enough to absorb those errors without losing reliability (Bekker & Goldberger, 2016).

4.2. *Symptom representation*

Symptoms and signs in plant leaves may come from a wide variety of sources, which include diseases, nutritional deficiencies, pests, phytotoxicity, mechanical damage, cold and heat damage, etc. As a result, a truly comprehensive diagnosis system should be able to deal with a classification problem having many classes. Although deep learning techniques have a remarkable ability to classify a large number of classes, currently there are not enough data available to make such a comprehensive diagnosis system feasible. In practice, only a few more common and relevant diseases are usually considered. As a consequence, when tools for disease recognition are used under real world conditions, they have to find the class that best explains the symptoms among the limited subset of disorders for which they were trained, often leading to incorrect diagnosis.

The option to consider only the most relevant disorders when training the classifier may seem logical, as this will cover the majority of practical situations. This is even more evident given the difficulty of obtaining a sufficient number samples for lesser common disorders. Nevertheless, such an option brings an undesirable consequence. Diseases that occur more frequently are, in general, better known and, as a consequence, more easily recognised by farmers and farmworkers. Rarer disorders have a higher probability of being misidentified and mishandled. Fortunately, in most cases rare diseases tend to have low impact on crop production. On the other hand, misclassification may result in the application of pesticides that will be ineffective against that disorder, unnecessarily increasing costs.

4.3. *Covariate shift*

Covariate shift is the phenomenon in which differences between the distributions of the training data used to train the model and the data on which the model is to be applied result in low accuracies (Barbedo, 2016; Sugiyama, Nakajima, Kashima, Bünau, & Kawanabe, 2007). There is an entire area, the so-called "domain adaptation", dedicated to mitigating this problem (Ben-David et al., 2010). This problem becomes evident when the same database is used for training and assessing the models, which is common practice in the machine learning community for practical reasons (Ferentinos, 2018). This author observed a very significant drop in accuracy when the model was tested with a limited amount of samples from other data sources. This confirmed the remarks previously made by Mohanty et al. (2016), who observed that when the model trained using the PlantVillage database was applied to images originating from trusted online sources, the accuracy quickly fell below 50%.

Again, the solution for this problem would be collecting a much wider variety of training data, which should include different geographic areas, cultivation conditions, and image circumstances (Ferentinos, 2018). As commented before, this is not an easy task, although the application of social network concepts may make such a task more feasible.

4.4. *Image background*

Traditional machine learning methods may be adversely affected by the image background, especially if it contains other leaves or soil (Barbedo, 2016). Because of that, leaf segmentation is sometimes an unavoidable step, and placing some kind of panel behind the leaf is a common requirement during image capture. Deep neural nets are known for learning the objects of interest even in busy images (Krizhevsky et al., 2012), so in theory the requirement for leaf segmentation could be relaxed. Indeed, experiments conducted by Mohanty et al. (2016) indicated that better results could be achieved by keeping the background intact. However, the images used in those experiments were collected using a regularised process that generated relatively homogeneous backgrounds. The experiments conducted in the context of the present work used images that, in some cases, had very busy backgrounds (Fig. 2) that produced slightly different results: the accuracy was around 76% for the original images and 79% for the images with the background removed. This is likely to be due to the fact that the backgrounds have elements that mimic the characteristics of certain diseases, so the network ends up learning them as well, leading to error. This seems to indicate that, if the background is busy and has elements that share characteristics with leaves and symptoms, removal may be useful. On the other hand, since removing the background is labour intensive and the expected performance improvement is not very substantial, in most cases this step can be skipped.

Fig. 2 – Example of busy background with elements mimicking those found on leaves.

Leaf segmentation, and the resulting background removal, is a difficult problem by itself. A possible way to deal with the background in a practical tool would be to allow the user to select the region of interest prior to the classification, which is relatively simple given that almost all mobile devices come with the touchscreen technology. This seems to be a simpler option than requiring the user to place a panel behind the leaf.

It is worth mentioning that humans are remarkably effective at extracting patterns from images, being more effective dealing with complex scenes than any computer program. However, as discussed in the introduction, trained individuals are not always readily available, underlining the importance of reliable automatic tools.

4.5. *Image capture conditions*

Many image databases are built using images captured under relatively controlled conditions. However, any technique intended to be used in practice has to be prepared to deal with images captured by different devices and people, at different angles and light conditions, and under different environmental circumstances. No matter the type of machine learning technique being applied, the training database has to reflect this reality. In the case of deep learning, this means that an extremely large number of labelled images are needed to ensure that the variety of conditions expected in practice are included, at least to an acceptable extent. Simply put, there is no database that satisfies this condition – new databases built using social network concepts may achieve this in the future, but this is still a very challenging task. This is an important aspect that many authors either treat superficially or simply ignore when analysing the results of their experiments.

Experiments carried out by Ferentinos (2018) clearly show the deleterious effects of training CNNs with an insufficiently comprehensive dataset: the success rate fell from 99% to 68% when a net trained with field-condition images was used to identify laboratory-condition images, and to a mere 33% when those sets (field and laboratory) were reversed. They remarked that images captured under actual cultivation conditions are essential for the development of useful disease recognition tools.

It is important to highlight that some undesirable illumination effects may cause irreversible loss of information. One of the most damaging of such effects is specular lighting, which is a high intensity reflection that occurs at certain angles of view (Barbedo, 2016). Because useful information is absent in areas affected by this phenomenon, including samples with specular reflection in the training database is not helpful. Fortunately, specular reflections can be easily minimised by altering the angle of capture and/or the position of the leaf, although some degree of reflection will almost always occur.

4.6. *Symptom segmentation*

As commented before, one of the main advantages of the deep learning approach is that the symptoms do not have to be explicitly identified in the image. Although, in this context, the problem of symptom segmentation becomes irrelevant, it may be useful to isolate the region where the symptom is located, as it contains most of the relevant information. The accuracy achieved by the CNN trained with individual lesions and localised symptom regions (Section 2) was 87%, higher than the accuracy obtained using the original images (76%). In order to determine which factor was the main one responsible for this improvement, the size of the expanded training set was reduced to match that of the original dataset. This led the accuracy to fall to 81%. This indicates that the increasing the number of samples and limiting the region of interest to the area where the symptoms are located had roughly the same impact on the results. Another advantage of considering individual lesions is that the predicted classes can be tallied into an overall diagnosis for the plant being analysed, thus decreasing the impact of individual misclassifications.

A system relying on images of specific diseased areas would require its users to select the appropriate regions prior to the classification process. In most cases, this task is relatively straightforward, especially with touchscreen technology being now widespread. However, in cases which the disease causes generalised damage over the entire leaf, the user may have trouble selecting a suitable region. Some selection criteria for cases like this were presented in Section 2, but they are subjective and may lead to diverse selections depending on the user. The model has to be prepared to deal with those differences, especially if there are multiple diseases with similar characteristics. In any case, it is worth noting that the advantages obtained by focussing on the regions of interest tend to diminish as larger databases are built.

It is also worth noting that with the use of localised leaf regions, there are no cues (e.g. leaf shape) for the system to derive the plant species. However, such information will normally be known by the user and can be easily entered as input to the system.

4.7. *Symptom variations*

Although the visual manifestations of most diseases have specific characteristics, there is always some variation in colour, shape and size of symptoms. Symptom variation creates problems for image-based diagnostics using the visible spectrum to delineate ‘healthy’ or ‘diseased’ pixels (Barbedo, 2016). The stage of the disease (symptom severity) is arguably the most important source of variability, as symptoms may range from very mild and barely visible in the beginning, to causing widespread necrosis in the most advanced stages of infection. As a result, diseases may be easier or harder to discriminate depending on the stage of infection.

If all variations expected for a given symptom are included in the training dataset, deep learning-based tools can properly deal with the challenges caused by such diversity. In practice, it is hard to predict the full extent of symptom variability associated with a given disease, and even more difficult to obtain enough images to represent each one of those situations. A more realistic approach is to continuously add new image samples to the database, gradually expanding it to a truly comprehensive set, thus reducing the impact of this problem (Barbedo, 2016).

4.8. *Multiple simultaneous disorders*

Images are usually labelled with a single disease. However, other diseases, as well as other kinds of disorders such as nutritional deficiencies and pests, may manifest simultaneously. This is actually a very common occurrence, because as a plant immune system is weakened by an infection, other disorders can more easily move in (Barbedo, 2016). A possible way to deal with this situation would be to create mixed classes, which would include all possible combinations of diseases. This is not a good option because the number of classes would grow greatly, increasing the chance of misclassification. Also, since the proportion of symptoms associated with each disease would be different in each image, the intra-class variability would be impractically high.

A more suitable solution would be to consider individual lesions and localised symptom regions, as discussed in Section 4.6. In this case, since lesions are processed separately, it would be possible to identify multiple simultaneous disorders. This approach also has some limitations. If the delimitation of the region of interest is to be carried out by the system's users, they have to notice the presence of distinct symptoms and then submit each one to the system separately. If region delimitation is to be done automatically, errors associated with the process may propagate and generate wrong predictions. Additionally, if the symptoms cohabit the same space, the result may depart dramatically from the characteristics expected for the original diseases (Barbedo, 2016), inevitably leading to error. Currently, possible solutions for the complex problem of simultaneous disorders are still lacking.

4.9. *Disorders with similar symptoms*

The number of agents (diseases, nutritional deficiencies, pests, mechanical damage, etc.) capable of producing lesions and other symptoms in a single host species is very high (Section 4.2). A few can produce similar symptoms that cannot be properly differentiated even by experts in plant pathology. Thus, sometimes visual cues are simply not enough to resolve possible classification ambiguities. Even if an image is captured with nearly perfect quality, the symptom characteristics may be too generic to allow a definite diagnosis. Human experts usually try to explore other cues, such as current weather, historical disease data, and overall state of the plant, to draw accurate conclusions. This kind of side information could be incorporated into disease recognition systems to improve classification (Kamilaris & Prenafeta-Boldú, 2018). In fact, disease recognition systems could be coupled with plant disease prediction models to refine the results. Another way to resolve ambiguities would be to use other spectral bands, like infrared (Belin, Rousseau, Boureau, & Caffier, 2013; Dammera, Möller, Rodemann, & Heppner, 2011; Oberti et al., 2014). However, this may increase the costs associated with image capture, and most mobile telecommunication devices are not capable of capturing images in additional bands (Barbedo, 2016). Even if all actions to acquire more information are employed, there are cases for which ambiguities can only be resolved by laboratory analysis. Hence, it is currently unrealistic to expect an automatic disease recognition system to achieve perfect accuracy when used under field conditions.

4.10. *Other factors*

The overfitting problem of deep learning models appears when the model describes random noise or errors rather than the underlying relationship (Liu et al., 2018). There are a few ways to avoid overfitting. If transfer learning is being applied, freezing the first layers and retraining only the last few ones can prevent those layers from overfitting the new data set. Also, dataset augmentation operations such as image rotation, mirror symmetry, brightness adjustment, and PCA jittering increase the diversity of the training set and cause the model to generalise better (Liu et al., 2018). Mohanty et al. (2016) addressed the issue of overfitting by varying the train and test set ratio. They argued that if the degradation of performance is not significant when the size of the training dataset is decreased, this indicates that the model is not overfitting the data.

Another problem is that many systems rely solely on images of the upper side of the leaves (Kamilaris & Prenafeta-Boldú, 2018). While many diseases indeed manifest in that region, frequently they are better characterised by signs present in other parts of the plant. Thus, it would be useful to expand the tools to be able to deal with images of any part of the plant, despite the fact that this would make the construction of a comprehensive database even more challenging.

Many systems proposed in the literature are being trained and tested using the PlantVillage database (Amara et al., 2017; Brahimi et al., 2017; Cruz et al., 2017; Ferentinos, 2018; Mohanty et al., 2016). This is explained by the fact that this is a reasonably extensive database that was formerly made freely available. However, considering that most authors are using similar network architectures, the experiments end up being quite redundant. Indeed, the results reported in the literature are, for the most part, similar. While it is true that many of those works examine different aspects of the disease recognition problem, the fact is that not much new information is being produced on the subject. In order for this field of research to advance, new tests with more challenging datasets should be performed, otherwise there will be too much repetition without producing additional knowledge.

5. Conclusion

The application of deep learning concepts to plant pathology problems is growing quickly. While the results reported so far in the literature are very encouraging, many aspects that may affect the practical adoption of such technologies are not being thoroughly investigated. The objective of this work was to discuss those aspects, offering some guidance to further advance this important field of research.

Many issues discussed in this article can be overcome, or at least minimised, by the expansion of available data for training and testing. Data sharing can greatly contribute toward the creation of more comprehensive datasets. This was the main motivation for making the database used in this work freely available for the research community. Future efforts could also focus on creating mechanisms to facilitate and encourage the involvement of farmers and plant pathologists in the process of image collection and labelling. Using plant parts other than leaves is another possibility that deserves further investigation. It is also worth noting that, as effective as current CNN architectures have been when applied to plant disease recognition, it may be possible devise new leaner configurations better suited for this application.

It is clear that there are still many challenges to overcome, and some problems still do not have suitable solutions. This indicates that, currently, tools for automatic recognition of plant diseases, rather than offering a definite answer, can at most provide a very educated guess that will allow its users to take some action in a timely manner, especially when specialised technical assistance is not available. As technology evolves, some of those limitations may be overcome, but there is still much to be investigated.

Acknowledgements

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

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. (2017). Automatic image-based detection and recognition of plant diseases – a critical view. In *Proceedings of the Brazilian conference on agroinformatics* (pp. 69–77).

Bekker, A. J., & Goldberger, J. (2016). Training deep neural-networks based on unreliable labels. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2682–2686).

Belin, E., Rousseau, D., Boureau, T., & Caffier, V. (2013). Thermography versus chlorophyll fluorescence imaging for detection and quantification of apple scab. *Computers and Electronics in Agriculture, 90*, 159–163.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning, 79*(1–2), 151–175.

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In *Proceedings of* *the workshop on unsupervised and transfer learning* (Vol. 27, pp. 17–37).

Bock, C. H., Poole, G. H., Parker, P. E., & Gottwald, T. R. (2010). Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Critical Reviews in Plant Sciences, 29*, 59–107.

Brahimi, M., Boukhalfa, K., & Moussaoui, A. (2017). Deep learning for tomato diseases: Classification and symptoms visualization. *Applied Artificial Intelligence, 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. *Frontiers in Plant Science, 8*, 1741.

Dammera, K.-H., Möller, B., Rodemann, B., & Heppner, D. (2011). Detection of head blight (Fusarium ssp.) in winter wheat by color and multispectral image analyses. *Crop Protection, 30*, 420–428.

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*(11), 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.

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of the annual conference on neural information processing systems* (pp. 1106–1114).

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. *Frontiers in Plant Science, 7*. article 1419.

Oberti, R., Marchi, M., Tirelli, P., Calcante, A., Iriti, M., & Borghese, A. N. (2014). Automatic detection of powdery mildew on grapevine leaves by image analysis: Optimal view-angle range to increase the sensitivity. *Computers and Electronics in Agriculture, 104*, 1–8.

Oppenheim, D., & Shani, G. (2017). Potato disease classification using convolution neural networks. *Advances in Animal Biosciences: Precision Agriculture, 8*(2), 244–249.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.

Sugiyama, M., Nakajima, B., Kashima, H., Bünau, P., & Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of advances in neural information processing systems* (Vol. 20, pp. 1433–1440).