2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Classification of Apple Tree Disorders Using Convolutional Neural Networks

Lucas G. Nachtigall and Ricardo M. Araujo
Federal University of Pelotas
Pelotas, RS, Brazil
lucas.nachtigall@gmail.com
ricardo@inf.ufpel.edu.br

Gilmar R. Nachtigall
Embrapa Grape & Wine
Vacaria, RS, Brazil
gilmar.nachtigall@embrapa.br

***Abstract*—This paper studies the use of Convolutional Neural Networks to automatically detect and classify diseases, nutritional deficiencies and damage by herbicides on apple trees from images of their leaves. This task is fundamental to guarantee a high quality of the resulting yields and is currently largely performed by experts in the field, which can severely limit scale and add to costs. By using a novel data set containing labeled examples consisting of 2539 images from 6 known disorders, we show that trained Convolutional Neural Networks are able to match or outperform experts in this task, achieving a 97.3% accuracy on a hold-out set.**

I. Introduction

Apple fruits have an important role in worldwide nutrition and commerce with over 80 million tons consumed yearly, making it one of the most consumed fruits in the world. The rapid diagnosis of disturbances in apple trees, such as nutritional imbalance, diseases and plagues, is becoming increasingly important not only to improve productivity but also to help define environmentally friendly policies for the use of fertilizers and agrotoxins.

The diagnosis of apple cultures is most often conducted by visual inspection of leaves and fruits, typically on site by an expert. Laboratory analysis can be used in more complex or new cases. However, this approach bears a high cost, as it requires experts that demand specialized training. The need for experts not only limit scale but may also reduce effectiveness due to human errors since experts are often specialized in a few types of disorders.

In this paper we report on results of training and applying machine learning models to automatically classify common disorders directly from images of apple tree leaves (*Malus domestica Borkh*). Our hypothesis is that the current state-of-the-art models, in particular deep Convolutional Neural Networks, and related training algorithms are able to attain performance comparable to or better than human experts.

In order to do so, we built a novel data set of labeled images containing examples of five of the most common and important disorders affecting this culture [Valdebenito-Sanhueza *et al.*, 2008; Nachtigall *et al.*, 2004]: *Glomerella*, Scab, Potassium Deficiency, Magnesium Deficiency and Herbicide Damage. We train Convolutional Neural Networks (CNN) on this data set, as they are often considered the state-of-the-art in image classification [Krizhevsky *et al.*, 2012], providing extensive analysis of the classification results, comparing them to those provided by other algorithms and by experts. We show that the trained CNN is able to outperform experts, with a 97.3% accuracy.

II. Related Work

There are a few commercial systems in use to help with the diagnosis of disorders in cultures, although not restricted or applicable to apple trees. For instance, BASF's Digilab[1] and EMBRAPA's Virtual Diagnose[2] provide tools to work over digitized images of leaves and compare them to a database of known disturbances. The comparison is entirely manual. Simple printed guides containing photos and explanations on how to diagnose a wide range of issues are also widely in use [Valdebenito-Sanhueza *et al.*, 2008].

Rumpf *et al.* [2010] aimed to discriminate healthy from unhealthy sugar beet leaves, to differentiate between three types of diseases and to identify diseases even before specific symptoms became visible. The authors used Support Vector Machines and as input they used nine spectral vegetation indexes, as features, resulting in classification accuracies up to 97% when differentiating healthy from unhealthy, 86% when distinguishing between three diseases and between 65% and 90% for pre-symptomatic detection of diseases. Notably, this approach requires specialized hardware to obtain the spectral images and considerable feature selection by the authors.

Al-Hiary *et al.* [2011] attempted at classifying six diseases from leaf images. The authors used 32 samples for each of the six classes of leaves and a Multilayer Perceptron to perform the classification. Features were manually defined as 10 texture features extracted from the image, showing a accuracies between 83% and 94%.

More recently, deep networks are also being applied to this domain. Revathi and Hemalatha [2014] focused on cotton leaf spot diseases. The authors used a data set with 270 images divided into 6 disease classes. Again, features were manually defined, consisting of leaf edge, color and texture features. A Cross Information Gain Deep Forward Neural Network was used to perform the classification, resulting in an overall

[1] http://www.agro.basf.com.br
[2] http://www.diagnose.cnptia.embrapa.br/diagnose/

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.75

accuracy of 95%. Tan *et al.* [2015] uses synthetic infrared images of diseased and healthy melons to train a CNN, which is allowed to extract features automatically, resulting in an accuracy of up to 97.5% when classifying as healthy or not.

III. Methodology

Our methodology consists of building a data set containing labeled images of five types of disorders commonly affecting apple orchards. This data set was randomly partitioned into training, validation and test subsets. The training and validation subsets were used to train and optimize a Convolutional Neural Network and a Multilayer Perceptron (as a baseline). The test set was then used to assess the performance of the resulting classifiers and was also presented to experts for classification in order to allow for a comparison. In what follows, we detail each of these steps.

A. Dataset

The dataset was built by harvesting leaves from three species of apple trees (*Maxigala*, *Fuji Suprema* and Pink Lady) and photographing each leaf over a white background. Each leaf was then subjected to laboratory tests to properly identify the underlying disorder. The disorder was used to label the image. Harvesting occurred between January and April 2015 from orchards located in the southern part of Brazil, at Embrapa Uva e vinho - Estação Experimental de Fruticultura de Clima Temperado, located in Vacaria, RS, Brazil (28º30'49 "S, 50º52'58" W).

Healthy leaves and five disorders were selected among those collected, as they are the most prevalent in the region. Selected symptoms represent two damages caused by nutritional imbalances (deficiency of potassium and magnesium), two diseases damage (apple scab and *Glomerella* stains and damage caused by herbicide (glyphosate). Table I summarizes the data set.

The identification of the disorders was conducted by a group of professional agronomist researchers specialized in these symptoms and with ample experience in plant nutrition and plant pathology. In order to reduce errors and properly establish a ground-truth, three strategies were employed by the experts to properly diagnose each issue.

a) For symptoms caused by nutritional imbalances, samples of normal leaves and leaves with potassium and magnesium deficiency symptoms were selected, each consisting of 100 leaves. The samples then were forwarded to the laboratory for chemical analyzes in order to quantify the total concentrations of nutrients (potassium and magnesium). The analysis results show that samples with symptoms effectively represent the deficiencies of potassium and magnesium.

b) For symptoms caused by disease damage (apple tree scab and Glomerella's stains), leaf samples were selected with symptoms previously identified for the two diseases. These samples were incubated for multiplication of the causative agent (fungus) and after was performed the isolation of fungi and their characterization and identification using a microscope, allowing the proof of causal agents and their damage on apple tree leaves. These images were performed from samples which leaves presented the proposed symptoms, showing that samples with diseases symptoms effectively represent the selected diseases.

c) For symptoms caused by herbicide damage (*glyphosate*), it was decided to conduct chemical analysis in order to quantify the total concentrations of nutrients which could possibly cause confounding of symptoms in cases where the nutrients concentrations were below normal. This decision was due the fact that the analysis of the herbicide's active principle is difficult to characterize, once it is rapidly degraded on the plant after its absorption and origin of toxicity symptoms. Then, samples with herbicide damage symptoms were put through the same protocols for nutritional analysis, which results showed these samples did not effectively present any nutritional disorders. Since the samples of leaves with symptoms caused by the herbicide *glyphosate* (in three levels of severity) were not different from the nutrient concentrations of normal leaves, all nutrients are within the range considered as standards for apple orchards [Nachtigall *et al.*, 2004].

Table I
Number of leaves collected for each class.

Issue	Number of leaves collected
Potassium deficiency	341
Magnesium deficiency	355
Scab damage	391
Glomerella stain	558
Herbicide damage	325
Healthy Leaves	569

Figure 1. Example image of leaf infected by Glomerella.

We used a white background to photograph each leaf separately, as show in Figure 1. The same camera was used for all pictures at a resolution of 12MP. A few images which presented defects or were outside the capture standards used were discarded.

The complete data set is available at https://www.dropbox.com/s/b81z064ohynhlgn/DataSet-AppleLeaves.zip

B. Pre-processing, Training and Evaluation

All images were re-sized to a resolution of 256x256 pixels. In order to have a balanced data set, we randomly selected 290 examples (labeled images) for each of the five classes containing symptoms. The resulting 1450 images were divided into three subsets. A test subset (hold-out set) was created by randomly choosing 15 images of each class. The remaining examples were further partitioned in a training set (192 examples, 70%) and validation set (83 examples, 30%). The test set size was chosen so as to allow for a comparison with experts, as a large test set would become exhausting for experts to classify, possibly resulting in increased human error.

The tested learning algorithms were trained over the training set using different parameters and configurations and then applied to the validation set. The best performing parameters (over the validation set) of each algorithm were then trained using training and validation sets combined and applied to the test set. The test set was also shown to experts for classification, so that a direct comparison was possible. Hence, all results reported in this paper are for the test set.

Furthermore, in order to analyze the number of samples needed for a satisfactory classification, smaller training subsets were created, randomly selecting 5, 10, 20, 50, 100, 150, 200, and 250 samples from each class containing symptoms. These training subsets were then tested using the test subset previously created, without any changes in the network configuration after the validation process was finished.

As a final experiment, 275 randomly selected healthy leaves were added to the train subset and 15 added to the test subset in order to evaluate the network capacity to distinguish healthy leaves from the five chosen symptoms. This was conducted as a separate experiment, as the experts did not have access to healthy leaves.

We evaluated the results by calculating the overall accuracy of each classifier and analyzing the resulting confusion matrix along with recall, precision and kappa statistic [Landis and Koch, 1977].

C. Convolutional Neural Network

We used Caffe [Jia *et al.*, 2014] and DIGITS [NVIDIA, 2015] tools to help in building, training and testing the Convolutional Neural Networks. Multiple architectures were tested, ranging from shallow networks with 4 layers to deep networks using small (3x3) convolution filters as shown in [Simonyan and Zisserman, 2014].

The best results over the validation set were obtained by the AlexNet architecture [Krizhevsky *et al.*, 2012]. This network consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers followed by softmax and dropout regularization. The training parameters were as follows: batch size of 2; a step learning policy was added with a gamma of 0.2; maximum of 300 epochs.

This network was used to generate the final results over the hold-out test set, reported in the next section.

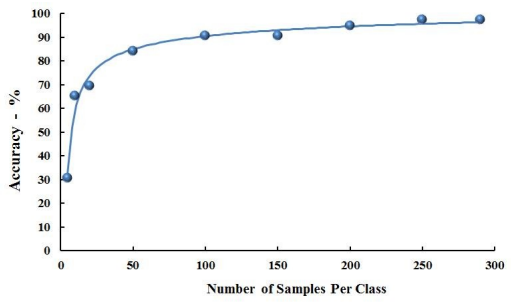

Figure 2. Relation between the number of samples used for learning and the accuracy obtained in the symptoms classification. The solid line represents a logarithmic best-fit function over the data points (circles).

D. Multilayer Perceptron

In order to provide a baseline comparison to CNNs, we applied Multilayer Perceptrons to the data set, using the same training methodology applied for CNNs. Multilayer Perceptron (MLP) was chosen for being closely related to CNNs. Shallow MLPs are able to achieve high accuracy on tasks such as digit recognition [Cho, 1997], image classification [Hara *et al.*, 1994] and feature extraction [Ruck *et al.*, 1990].

DIGITS was used to build and test the MLPs. Different shallow architectures were tested. The best configuration consisted of two hidden layers with 200 and 500 neurons, respectively, with one unit for each class at the output layer and 65536 input units (256x256 pixels).

E. Classification by Experts

To collect the classification by experts, we asked 7 volunteer researchers specialized in apple trees to classify the images in the test set, including the agronomist who collected the leaves for the database, since he had greater knowledge on the chosen disorders. These experts are further specialized in different fields of research, such as plant pathology or plant nutrition. Each expert was given a form to choose, for each image, one of the five classes. Healthy leaves were not shown to the experts.

Since experts are specialized in different disorders, in addition to a direct mean accuracy comparison we aggregated the experts' choices using a voting model to create a more reliable diagnosis. This is akin to collecting different opinions on particular symptoms, which is often conducted in practice. In this model, the diagnosis was given by the majority choice. Ties were broken by selecting the choice from the expert with the highest overall accuracy.

IV. Results

Figure 2 presents the overall CNN accuracy when different number of images per class are used for training and tested against the hold-out test set. The graph shows a logarithmic increase on accuracy when training set is increased. The curve largely levels off for samples larger than 200 images, evidencing that the number of samples collected was adequate.

Table II
CONFUSION MATRIX RESULTING FROM CNN CLASSIFICATION ON THE HOLD-OUT TEST SET.

		CNN						
Label Image	Symptom	Glomerella	Herbicide	Magnesium Def.	Potassium Def.	Scab	*Recall*	Precision
	Glomerella	15	0	0	0	0	100.0%	93.3%
	Herbicide	0	15	0	0	0	100.0%	100.0%
	Magnesium Def.	0	0	15	0	0	100.0%	100.0%
	Potassium Def.	0	0	0	14	1	93.3%	100.0%
	Scab	1	0	0	0	14	93.3%	93.3%
Accuracy								97.3%
Kappa								0.97

Table III
FIELD OF RESEARCH AND ACCURACY OBTAINED BY EACH EXPERT WHEN CLASSIFYING IMAGES IN THE HOLD-OUT TEST SET.

Subject	Field of research	Accuracy
1	Soil	93.3%
2	Plant pathology	92.0%
3	Post harvest	90.6%
4	Plant pathology	70.6%
5	Plant nutrition	60.0%
6	Crop Science	60.0%
6	Environmental management	37.3%
Average	-	71.9%

Table II shows the final CNN confusion matrix when applied to the hold-out test set. The overall accuracy was of 97.3%, with only two incorrect classifications. The MLP, applied to the same test set, resulted in an accuracy of 77.3%.

Table III shows the individual accuracy obtained from the 7 consulted experts, also showing their specific field of research. We can observe that accuracy varies considerably across experts. The best result (93.3%) is worse than the result obtained by the CNN, but the average (71.9%) was much worse and below that obtained by the MLP. Table IV presents the confusion matrix of the experts when aggregated by voting, where it can be seen that the overall accuracy improves significantly.

Table V summarizes the results. The best accuracy was obtained by the CNN, with a 97.3% accuracy, followed by the voting system with 96.0% accuracy, the best expert with 93.3% accuracy, and the MLP network with 77.3% accuracy. Figure 3 shows the Confidence interval ($IC_{1-\alpha}(p)$) for the classifiers, using a 99% confidence level. It is possible to observe that all techniques are much better than random choice and that aggregated experts and the CNN have comparable performance and both are better than the MLP.

A second experiment was conducted, introducing healthy leaves to the data set. In this case, only the CNN was tested. The same distribution of 275 images for training and 15 images for testing was used.

Table IV
CONFUSION MATRIX RESULTING FROM THE AGGREGATION OF THE CLASSIFICATIONS PROVIDED BY HUMAN EXPERTS.

		Voting System						
Label Image	Symptom	Glomerella	Herbicide	Magnesium Def.	Potassium Def.	Scab	*Recall*	Precision
	Glomerella	15	0	0	0	0	100.0%	100.0%
	Herbicide	0	14	1	0	0	93.3%	100.0%
	Magnesium Def.	0	0	14	0	1	93.3%	93.7%
	Potassium Def.	0	0	0	14	1	93.3%	100.0%
	Scab	0	0	0	0	15	100.0%	88.2%
Accuracy								96.0%
Kappa								0.95

Table V
SUMMARY OF THE RESULTS, ORDERED BY ACCURACY.

Technique	Accuracy
CNN	97.3%
MLP	77.3%
Voting system	96.0%
Highest Accuracy Expert	93.3%

No changes were made in the configuration of the CNN network in order to avoid an over fitting to the results. With healthy leaves in the training and test groups, the CNN was able to achieve accuracy of 96.67%, as shown in Table VI. It is possible to observe that the trained CNN is able to attain perfect accuracy when distinguishing between healthy and unhealthy leaves. This is expected, as the disorders all display strong symptoms on the leaves, but an additional classification error is now made when distinguishing between disorders.

One possibility to improve classification is to train a CNN to first distinguish between healthy and unhealthy and then another to further distinguish between disorders. Indeed, a CNN trained on a binary healthy/unhealthy class is also able to attain 100% accuracy, hence allowing the use of the disorders-only CNN.

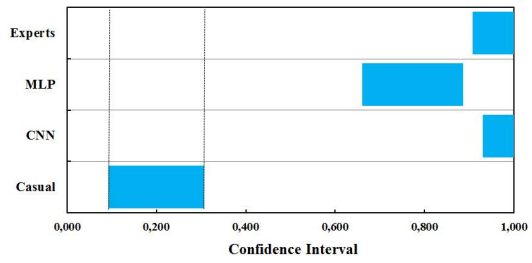

Figure 3. Confidence intervals with 99% confidence level, with normal approximation for each share of accuracy.

Table VI
CONFUSION MATRIX RESULTING FROM CNN CLASSIFICATIONS ON A HOLD-OUT TEST SET WHEN HEALTHY LEAVES ARE ADDED TO THE DATA SET.

		CNN							
Label Image	Symptom	Glomerella	Herbicide	Magnesium Def.	Potassium Def.	Scab	Healthy Leaves	*Recall*	Precision
	Glomerella	15	0	0	0	0	0	100.0%	93.7%
	Herbicide	0	15	0	0	0	0	100.0%	100.0%
	Magnesium Def.	0	0	15	0	0	0	100.0%	93.7%
	Potassium Def.	0	0	0	14	1	0	93.3%	100.0%
	Scab	1	0	1	0	13	0	86.6%	92.8%
	Healthy Leaves	0	0	0	0	0	15	100.0%	100.0%
Accuracy									96.6%
Kappa									0.96

V. CONCLUSIONS AND FUTURE WORK

Our results show that a CNN based on the AlexNet architecture is able to significantly outperform the baseline MLP, showing comparable performance to that of a group experts and outperforming any single expert. Moreover, perfect accuracy was obtained when only distinguishing between healthy and unhealthy leaves.

We conclude that CNNs compose a viable and useful option for this task, with more robust classifications than single human experts. In this sense, an automated system based on the trained model could contribute towards diagnosis reliability and cost reduction.

Compared to previous works, our approach does not require specialized equipment to capture the images or any sort of feature extraction or engineering. The CNN is able to learn relevant features from the data, to which we attribute the improved performance. This also allows for the general approach to be used in different disorders or even cultures with changes only to the data set. This is important to allow for the automatic improvement of the model when more data is made available.

Although the SVM technique showed high accuracies in the related work, when applied the same methodology as the other techniques, using no pre-processing or feature extraction, the results did not achieve more than 60% accuracies, therefore they were not included in this article.

Several lines of future work are being planned. We are expanding the current data set to make available more diverse examples. While we have shown that more examples obtained in the same way will only provide marginal improvements, the introduction of more diversity (e.g. different backgrounds and light conditions) could allow for better performance. We are also introducing additional disorders and cultures to test how the approach scales with these settings.

Different architectures are also being considered. We believe that a combination of more examples and improved architecture could lead to a system that can consistently outperform experts. Finally, we aim at integrating our methodology into working systems that can be used on the field, in less controlled conditions.

REFERENCES

H Al-Hiary, S Bani-Ahmad, M Reyalat, M Braik, and Z AL-Rahamneh. Fast and accurate detection and classification of plant diseases. *Machine learning*, 14:5, 2011.

Sung-Bae Cho. Neural-network classifiers for recognizing totally unconstrained handwritten numerals. *Neural Networks, IEEE Transactions on*, 8(1):43–53, 1997.

Yoshihisa Hara, Robert G Atkins, Simon H Yueh, Robert T Shin, and Jin Au Kong. Application of neural networks to radar image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 32(1):100–109, 1994.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

Gilmar Ribeiro Nachtigall, C Basso, and C J S Freire. Nutrição e adubação de pomares. In *NACHTIGALL, G.R. (Ed.) Maçã:*, pages 66–77. Embrapa Informação Tecnológica, Produção. Bento Gonçalves: Embrapa Uva e Vinho; Brasília, 2004.

NVIDIA. Nvidia digits – interactive deep learning gpu training system. https://github.com/NVIDIA/DIGITS, 2015. Accessed: 2016-02-02.

P Revathi and M Hemalatha. Identification of cotton diseases based on cross information gain_deep forward neural network classifier with pso feature selection. *International Journal of Engineering and Technology (IJET) ISSN*, 2014.

Dennis W Ruck, Steven K Rogers, and Matthew Kabrisky. Feature selection using a multilayer perceptron. *Journal of Neural Network Computing*, 2(2):40–48, 1990.

T Rumpf, A K Mahlein, U Steiner, E C Oerke, H W Dehne, and L Plümer. Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, 74(1):91–99, 2010.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Wenxue Tan, Chunjiang Zhao, and Huarui Wu. Intelligent alerting for fruit-melon lesion image based on momentum deep learning. *Multimedia Tools and Applications*, pages 1–21, 2015.

RM Valdebenito-Sanhueza, GR Nachtigall, A Kovaleski, RS dos Santos, and P Spolti. *Manual de identificação e controle de doenças, pragas e desequilíbrios nutricionais da macieira*. Embrapa Uva e Vinho, Bento Gonçalves, 2008.