Automatic Classification of Soybean Diseases Based on Digital Images of Leaf Symptoms

Jayme Garcia Arnal Barbedo[1], *Cláudia Vieira Godoy*[2]

[1] Embrapa Informática Agropecuária, Campinas, São Paulo, Brasil,
jayme.barbedo@embrapa.br

[2] Embrapa Soja, Londrina, Paraná, Brasil
claudia.godoy@embrapa.br

ABSTRACT

This paper presents an algorithm for automatic classification of diseases that produce symptoms in soybean leaves. The algorithm is based on digital image processing techniques and on a modified pairwise voting system that yields, at its output, a list of diseases with the respective likelihoods of being present in that leaf. Only color information is used, which is done by transforming the original RGB format into the HSV, L*a*b* and CMYK color spaces, and then extracting the intensity histograms from the grayscale representations of each one of the ten resulting channels. The capabilities of the algorithm were stressed by considering nine different diseases, and the results revealed that most diseases can be distinguished, however in some cases the symptoms are so closely related that information other than visual may be necessary for a reliable estimation.

KEYWORDS: disease classification, soybean leaves, digital image processing, color transformation

RESUMO

Este artigo apresenta um algoritmo para classificação automática de doenças que produzem sintomas em folhas de soja. O algoritmo é baseado em técnicas de processamento digital de imagens e em um sistema de votação por pares que produz, em sua saída, uma lista de doenças com as respectivas probabilidades de estarem presentes naquela folha. Apenas informação de cor é usada, o que é feito transformando o formato RGB original nos espaços de cor HSV, L*a*b* e CMYK, e então extraindo os histogramas de intensidade das representações em escala de cinza de cada um dos dez canais resultantes. As capacidades do algoritmo foram testadas a fundo pela inclusão de nove doenças diferentes, e os resultados revelaram que a maior parte das doenças pode ser distinguida, porém em alguns casos os sintomas são tão similares que informações além das visuais podem ser necessárias para uma estimativa confiável.

PALAVRAS-CHAVE: classificação de doenças, folhas de soja, processamento digital de imagens, transformação de cor.

INTRODUCTION

The detection of diseases in plants has been traditionally carried out by visual detection of symptoms in the field. The identification (classification) of those diseases is also usually manual, either by analyzing the characteristics of the symptoms, or by performing some laboratorial analysis. Despite the steady growth of automation in agriculture, phytopathology processes remain largely unchanged. Although some activities will always depend on agricultural engineers and phytopathologists to be carried out properly, many losses could be prevented and much faster responses could be achieved if reliable computer-aided detection and classification of diseases was available.

Many automatic methods for detecting and identifying diseases were proposed so far in the literature (BARBEDO, 2013), however most of them are very limited in their scope, being able to discriminate only a few diseases, such as those proposed by Pydipati, Burks and Lee (2006), Sanyal and Patel (2008) and Phadikar, Sil and Das (2013). In practice, however, the number of diseases and other problems, such as nutritional deficiencies and pests, is much higher, and there may be a lot of confusion regarding the visual cues provided by the symptoms. This fact motivated the development of a strategy that was specifically designed to deal with a large number of diseases and other disorders (BARBEDO AND COSTA, 2015). This computer-aided disease identification method was based on a pairwise classification, combined with a modified voting system that outputs a lists of diseases with the respective probability, and it was developed and validated using maize leaf images.

The research presented in this paper took the ideas presented in Barbedo and Costa (2015) and repurposed them for the classification of nine different soybean diseases, always having digital images of symptomatic leaves as the sole source of information. The original algorithm was improved by fully automating the segmentation of the symptoms, which was originally done manually. The results confirmed that the method can be applied to plant species other than maize, and that the automation of the segmentation process did not have adverse effects over the method's performance. Additionally, the proposed method is easy to implement, and its modularity makes it straightforward to include more diseases and to retrain certain parts of the algorithm without the need to go through the complete training process again.

MATERIAL AND METHODS

Image Dataset

The dataset used in this work was composed by 372 images of soybean leaves containing symptoms of 9 different diseases. Approximately 70% of the images were used for training and tuning the algorithm, and the remainder ones were used in the tests, as shown in Table 1.

Table 1 - Image dataset composition.

Disease	Training	Tests	Total
Bacterial Blight	38	18	56
Rust	45	20	65
Phytotoxicity	16	7	23
Stem Canker	15	7	22
Corynespora Leaf Spot	43	19	62
Myrothecium Leaf Blight	1	1	2
Downy Mildew	31	15	46
Powdery Mildew	52	24	76
Septoria Brown Spot	14	6	20
All	*255*	*117*	*372*

Most images were captured under controlled conditions, however about 10% of them (35) were captured in the field, under different lighting conditions, in order to test the algorithm with situations that were not contemplated in the training.

The Algorithm

As commented before, the algorithm used in this work is roughly the same as the one used in Barbedo and Costa (2015) for maize leaves, with the only difference being found in the second box of Figure 1: here, the symptom segmentation is performed automatically, while in the original algorithm this step was manual.

Basic Processing

The first step of the algorithm is the segmentation of the symptoms. It was determined empirically that the best color channel for this purpose is *H*, from the HSV (Hue, Saturation, Value) color space. The grayscale representation of this channel is taken, and the pixel values are rescaled so the value of the darkest pixel is zero, and the value of the brightest pixel is one. Then, a threshold is applied so that only those pixels with values smaller than 0.31 in this channel are kept in the original RGB (Red, Green, Blue) image. This value was chosen because, considering the image database used in this work, all symptoms were always considered in their entirety. The disadvantage of using such a fixed threshold is that parts of the healthy tissue will also be considered in most cases. This has a limited impact on the algorithm performance though, because the histogram-based procedure described later in this section is capable of at least partially compensating for this problem.

Fig. 1 - Basic structure of the algorithm.

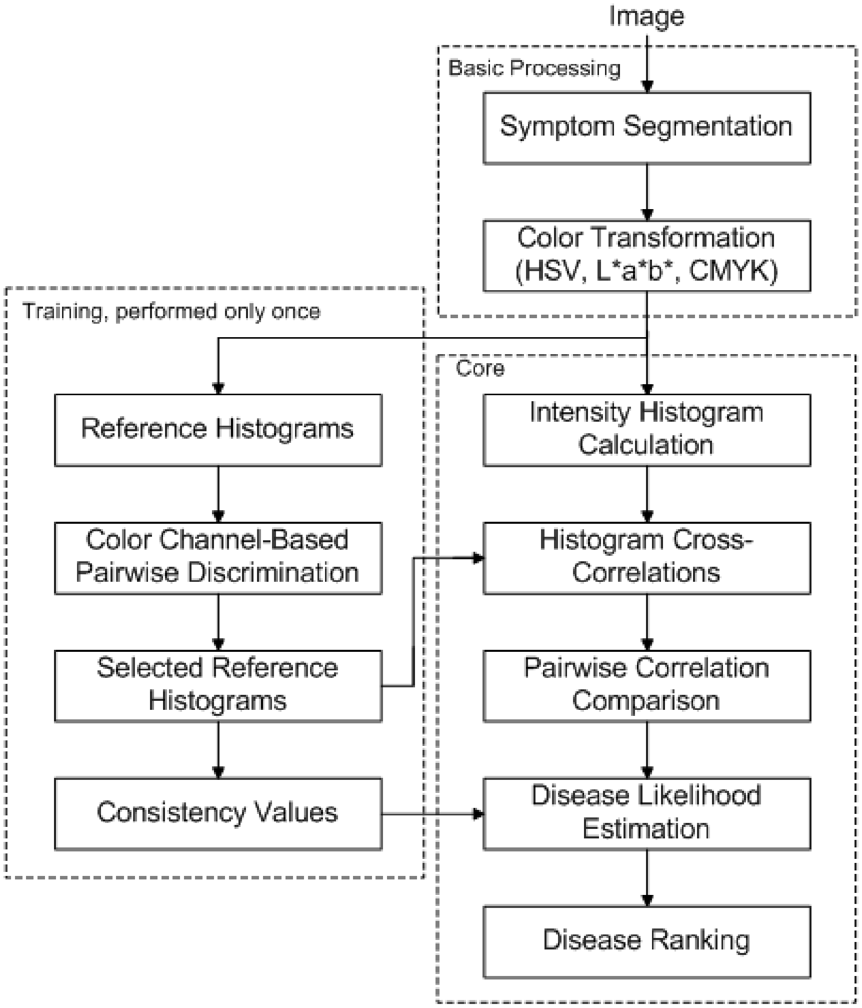


Image
Basic Processing
Symptom Segmentation
Color Transformation (HSV, L*a*b*, CMYK)
Training, performed only once
Core
Reference Histograms
Color Channel-Based Pairwise Discrimination
Selected Reference Histograms
Consistency Values
Intensity Histogram Calculation
Histogram Cross-Correlations
Pairwise Correlation Comparison
Disease Likelihood Estimation
Disease Ranking

Source: Barbedo and Costa (2015)

After the unsuitable pixels are blackled out, the original image is transformed from the original RGB format to the HSV, L*a*b* (Lighness and color opponent components) and CMYK (Cyan, Magenta, Yellow and Key).

Training

The training stage was ran only once, and is only necessary if the algorithm is to be updated with new images and/or new diseases. The first step of the training was the calculation of the reference histograms. This was done by taking each image of each disease in the training set and determining one 100-bin intensity histogram for the each color channel. Thus, for each image in the training set, ten histograms were generated. After that, all images corresponding to the same disease were combined by summing the respective histograms, thus generating ten reference histograms (one for each color channel) for each disease.

One way to deal with classification problems whose classes may not be well defined, as is the case here, is to divide the problem containing *c* classes into c(c-1)/2 two-class problems, an approach known as pairwise classification (Park and Fürnkranz, 2007). Here, since nine diseases were considered, there were 36 pairs of diseases. At this point, the color channels whose reference histograms correlated the least for each pair of diseases were taken as the ones with the best discriminative capabilities for those pairs. Therefore, each pair of diseases have two histograms associated, which are the histograms of the selected color channel that correspond to the two diseases in that pair.

Finally, the consistency values were calculated. For that, the cross-correlations between each selected reference histogram and the histograms of all corresponding images in the training set were calculated and averaged. The closer the resulting value was to one, the more consistent was the color channel for that disease, and hence the stronger the results based on it. Since there were 36 pairs of diseases, and the calculations were performed for both diseases in each pair, 72 consistency values were stored.

Core

After an image goes through the basic part of the algorithm, the intensity histograms for all ten resulting channels are calculated. In the following, the cross-correlations $X_{c,d}$ between those intensity histograms and the reference ones are calculated, where c is the color channel and d is the disease.

In the next step, each pair of diseases is analyzed as an independent problem. For each pair, the two corresponding cross-correlations $X_{c,d}$ are selected, where c is the selected color channel for that pair and d corresponds to the two diseases in that pair. The correlation differences are then calculated according to

$$CD_{d_1} = X_{c_{d1,d2},d_1} - X_{c_{d1,d2},d_2} \tag{1}$$

$$CD_{d_2} = X_{c_{d1,d2},d_2} - X_{c_{d1,d2},d_1} \tag{2}$$

where d_1 and d_2 are the first and second disease in the pair, respectively, and $c_{d1,d2}$ is the color channel corresponding to the (d_1,d_2) disease pair. The larger is the correlation difference *CD* for a given disease, the stronger is the indication that the symptoms are more closely related to such disease, and vice versa. CD_{d1} and CD_{d2} are then stored in the correlation difference vectors v_1 and v_2. The same procedure is repeated for all pairs of diseases, so the vector corresponding to each disease will have nine correlation difference values.

The next step is the calculation of the likelihood that the symptoms were produced by each of the diseases, according to:

$$L_d = \frac{\sum_{i=D,i\neq d}\ (v_{d,i} \cdot c_{d,i})}{\sum_{i=D,i\neq d}\ (c_{d,i})}, \qquad (3)$$

where L is the likelihood, d indicates the current disease, D is the set of all diseases, and c are the consistency values calculated in the training part. The index (d,i) indicates that the value corresponds to the pair containing the current disease d and disease i, with $i \in D$. Finally, all diseases are ranked from the highest to the lowest likelihood.

RESULTS AND DISCUSSION

Table 2 shows the percentage of times each disease appears in each position of the disease ranking returned by the algorithm.

Table 2 - Percentage of times each disease appears in each position of the ranking.

Correct Disease	**Position**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Bacterial Blight	37.5	55.4	1.8	0.0	1.8	1.8	1.8	0.0	0.0
Rust	74.6	3.4	0.0	0.0	3.4	11.9	6.8	0.0	0.0
Phytotoxicity	73.9	21.7	4.3	0.0	0.0	0.0	0.0	0.0	0.0
Stem Canker	9.1	22.7	27.3	22.7	13.6	4.5	0.0	0.0	0.0
Corynespora Leaf Spot	19.4	41.9	21.0	9.7	4.8	3.2	0.0	0.0	0.0
Myrothecium Leaf Blight	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Downy Mildew	91.3	8.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Powdery Mildew	61.8	25.0	1.3	0.0	0.0	0.0	0.0	3.9	7.9
Septoria Brown Spot	20.0	5.0	20.0	20.0	25.0	10.0	0.0	0.0	0.0
All	*54.2*	*20.4*	*8.4*	*5.8*	*5.4*	*3.5*	*1.0*	*0.4*	*0.8*

Table 3 shows the resulting confusion matrix if an absolute approach was adopted and the first ranked disease was taken as the algorithm's final diagnosis.

As it can be seen in Table 3, some diseases, like Myrothecium Leaf Blight and Downy Mildew, are almost always unequivocally correctly identified as the one that produced the corresponding symptoms. On the other hand, diseases like Stem Canker and Septoria Brown Spot are often confounded with other ones. There are three main reasons for those errors:

Table 3 - Confusion matrix considering the first ranked disease as the algorithm's diagnosis.

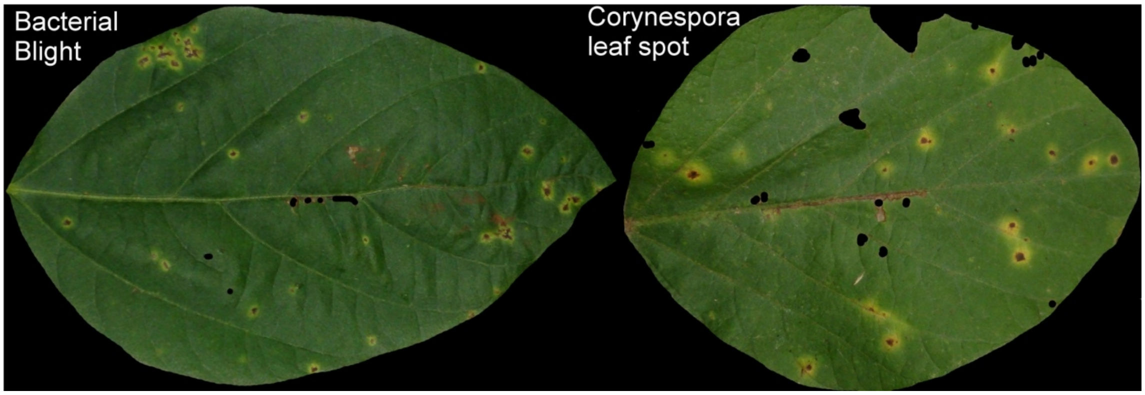
	Bacterial Blight	Rust	Phytoto-xicity	Stem Canker	Coryn. Leaf Spot	Myrot. leaf blight	Downy Mildew	Powdery Mildew	Septoria brown spot
Bacterial Blight	37.5	0.0	48.2	0.0	5.4	5.4	1.8	0.0	1.8
Rust	8.1	71.0	0.0	3.2	8.1	0.0	0.0	0.0	9.7
Phytoto-xicity	13.0	0.0	73.9	0.0	0.0	0.0	0.0	0.0	13.0
Stem Canker	40.9	9.1	9.1	9.1	13.6	0.0	4.5	0.0	13.6
Coryn. Leaf Spot	56.5	0.0	4.8	1.6	19.4	0.0	6.5	0.0	11.3
Myrot. leaf blight	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
Downy Mildew	0.0	0.0	0.0	0.0	0.0	0.0	91.3	8.7	0.0
Powdery Mildew	10.5	0.0	1.3	0.0	0.0	0.0	26.3	61.8	0.0
Septoria brown spot	30.0	5.0	10.0	0.0	10.0	0.0	25.0	0.0	20.0

1) Similarity of symptoms between diseases: some diseases produce visually similar symptoms. This is the case of the example shown in Figure 2 (bacterial blight versus Corynespora leaf spot), in which the differences are so subtle that even experienced technicians may have problems identifying the diseases correctly. In cases like this, it may be necessary to either couple the digital image processing-based algorithm with some other tool, such as an Expert System (JACKSON, 1999), or resort to external results, such as laboratorial analysis, in order to resolve the ambiguity.

2) Diseases with significant symptom variations: sometimes a disease may produce quite different symptoms depending on the stage of its life cycle, position on the leaf, and interaction with a variety environmental variables. Figure 3 shows an example of symptom variation for the bacterial blight disease. A way to deal with this problem is to identify all possible variations and populate the database with a large number of samples for all those variations. This is a very difficult task, as it is highly dependent on the existence of the right conditions, on the opportunity for the image capture being identified by an expert, and on the availability of personnel to capture the images under the right conditions.

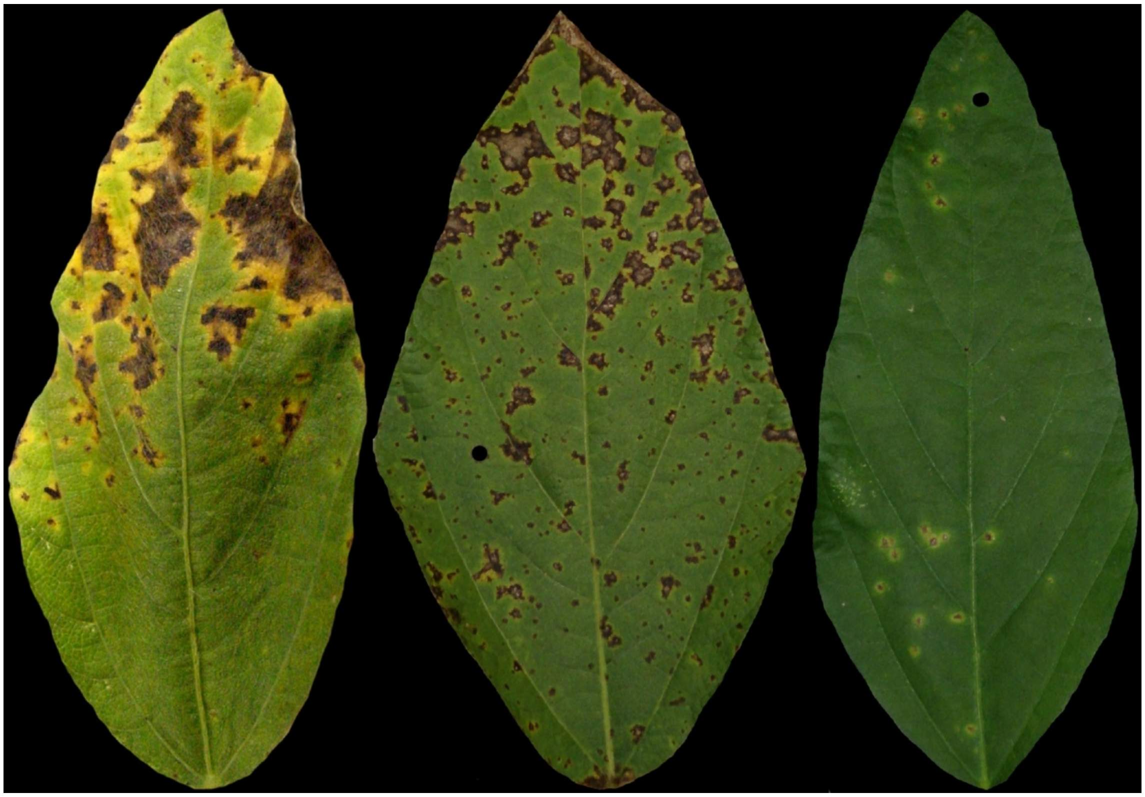
3) Conditions variations: the conditions under which the images were captured may have a strong influence on the results. Some color channels may partially compensate for illumination differences, however of those are too marked, the algorithm will have problems. Figure 4 shows an example of condition variations for two images showing symptoms of powdery mildew.

Fig. 2 - Example of different diseases producing similar symptoms.



Source: authors.

Fig. 3 - Example of a disease (bacterial blight) producing very different symptoms.

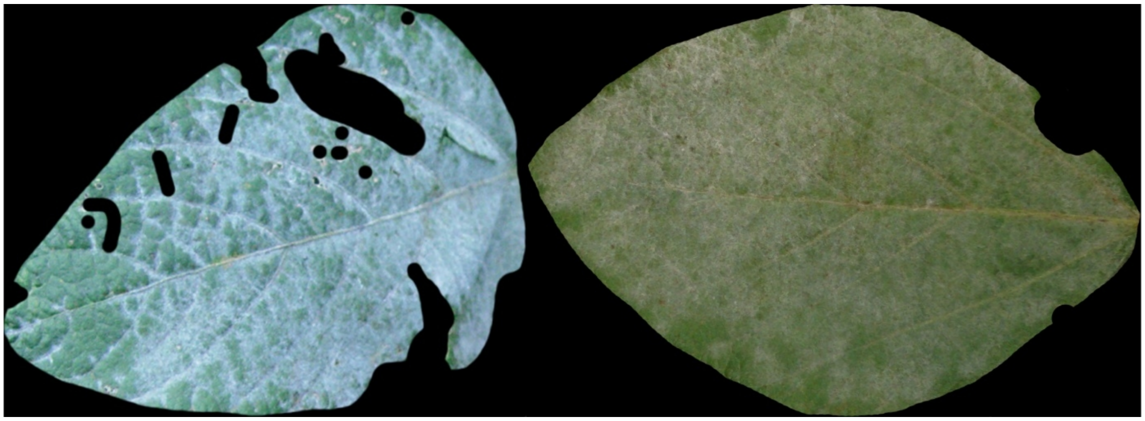


Source: authors.

Other factors, like specular light and shadows cast over the leaves may also cause the algorithm to misestimate the likelihoods for the diseases.

These results are a strong indication that, although digital image processing-based methods can be very useful to aid in the diagnosis of diseases, when more complex scenarios (such as considering several diseases) are tackled, the use of other sources of information may be unavoidable if unambiguous answers are expected. This also explains why all methods proposed in the literature so far only were successful when they tackled just a few diseases with highly dissimilar characteristics. Because there are no methods in the literature designed to identify a large number of diseases, no tests comparing different algorithms is presented here. As a final remark, it seems unlikely that computer-based systems will ever be able to completely replace plant pathologists and other specialists in plant sciences, even if the best tools from different disciplines are combined into a single powerful system. This is so because there are so many particularities and variables that interfere with the diagnosis process, that is impractical, at least in the near future, to fully take into consideration all those variations, in which case the creativity and flexibility of human assessments are invaluable. On the other hand, computer aided systems may provide valuable information for a quick first response, may monitor vast areas that would be inaccessible otherwise, and may very useful for farmers that do not have access to agricultural engineers, which is still the case for many people around the world.

Fig. 4 - Example of a images of symptoms of a same disease captured under different conditions.



Source: authors.

CONCLUSIONS

This paper presented the application of a digital image processing-based algorithm for identification of diseases in soybean plants. The original algorithm was developed for maize, and uses images of leaves to calculate the likelihood that the symptoms that are visible in the surface of the leaves were produced by each of the diseases considered during the algorithm's development. The algorithm applied to soybean leaves was identical to the original, with the exception that here the symptom segmentation was performed automatically. The results have shown that some diseases are very successfully identified, while others may have characteristics that make an unambiguous identification very difficult if other kinds of information external to the algorithm are not included. The problem of plant disease diagnosis is of such a complexity that it is unlikely that unambiguous answers will ever be possible without human involvement. However, computer-aided tools may be very useful in many situations, and there is still much room for improvement. Future research will concentrate on five fronts: a) improving the symptom segmentation by including an adaptive threshold; b) better exploring the differences between pairs of diseases; c) expanding the image database to include both more diseases and more samples of the diseases already considered; d) extending the algorithm to other plant species; e) coupling the proposed digital image-based algorithm with a expert system capable of resolving some of the ambiguities observed in the tests. The database and the latest implementation of the algorithm will be made available at https://www.agropediabrasilis.cnptia.embrapa.br/web/digipathos as soon as copyright and license issues are resolved.

ACKNOWLEDGEMENTS

The authors would like to thank Fapesp (proc. 2013/06884-8) and Embrapa (SEG 03.13.00.062.00.00 and SEG 02.14.09.001.00.00) for funding.

REFERENCES

BARBEDO, J. G. A. Digital image processing techniques for detecting, quantifying and classifying plant diseases. SpringerPlus, v. 2, n. 660, 2013.

BARBEDO, J. G. A.; COSTA, R. V. Identifying Multiple Plant Diseases Using Digital Image Processing. Biosystems Engineering, submitted in 2015.

PARK, S.-H.; FÜRNKRANZ, J. Efficient Pairwise Classification. In: EUROPEAN CONFERENCE ON MACHINE LEARNING, 2006, Berlin. Proceedings... 2006, p. 658–665.

PHADIKAR, S.; SIL, J.; DAS, A. K. Rice diseases classification using feature selection and rule generation techniques. Computers and Electronics in Agriculture, v. 90, p. 76–85, 2013.

PYDIPATI, R.; BURKS, T. F.; LEE, W. S. Identification of citrus disease using color texture features and discriminant analysis. Computers and Electronics in Agriculture, v. 52, n.1–2, p. 49–59, 2006.

SANYAL, P.; PATEL, S. C. Pattern recognition method to detect two diseases in rice plants. Imaging Science Journal, v. 56, n. 6, p. 319–325, 2006.