
www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(2), April 2025, 176-182. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i2.1079

Color Segmentation and Machine Learning for Disease Recognition in Rice Crop Leaves

Rocío Ochoa-Montiel, Carlos Sánchez-López, Fredy Montalvo-Galicia
Autonomous University of Tlaxcala, Faculty of Basic Sciences, Engineering and Technology. Tlaxcala, México.
E-mails: {1ma.rocio.ochoa*,1mongafre}@gmail.com, 1carlsanmx@yahoo.com.mx

Abstract. Timely detection of diseases in various crops is a necessary task to ensure sufficient production of food sources. Visual analysis by an expert is the method traditionally used for this activity, so it is subjective and prone to errors. In this paper, we propose a color segmentation method and a feature analysis for the recognition of rice crop leaves using machine learning. We use balanced sets of images and propose a set of experiments that allow us to discover the features that influence the classification indices, like the need to identify more precise characteristics for the classes of similar leaves or the disease regions. Results show that some features of texture and color are irrelevant for disease recognition.
Keywords: Color segmentation, Plant diseases, Machine learning.

Article Info
Received Jan 26, 2025
Accepted Mar 11, 2025

1 Introduction

Agriculture is an activity of vital importance for the economic development of a country since a constant source of food results in food security for the population and impacts its productivity [17]. The promotion of sustainable agricultural systems involves promoting the use of appropriate technologies to support the development of agriculture [16]. In this regard, computer vision techniques are a feasible alternative to address problems such as the timely detection of plant diseases because it is a task typically performed by human inspection, causing subjectivity in the visual analysis [8]. Several proposals have addressed this problem, using strategies based on classical, automatic, or hybrid learning. In the first case, it is common to use image processing and machine learning techniques [15, 6]. On the other hand, automatic learning strategies are models composed of several processing levels, such as convolutional neural networks (CNNs), symbolic learning models, and generative neural networks, among others [13, 14]. In hybrid strategies, it is common to use CNNs as feature extractors and classical machine learning techniques for classification [4].

Currently, the strategic crops for survival considered by FAO [7] are maize, beans, rice and wheat, among others. In countries such as Mexico, these grains are considered basic to guarantee food security [5]. In this work, we focus on the recognition of diseases in the leaves of rice crops using color segmentation and machine learning, since in the found literature there are limitations regarding the characteristics of the datasets and the classification strategies [12, 9, 1, 2, 8, 18-26]. To address this problem, we propose a color segmentation method, and a comparative analysis of features for disease recognition in rice leaves using machine learning. The diseases addressed in this work are Brown spot, Leaf blast, and Hispa.

In the next section, the theoretical basis is described. Section 3 presents the proposed methodology. Experiments and results description is in Section 4. Conclusions are in section 5.

2 Theoretical bases

Visual inspection is the most frequent method for disease detection in different crops. Proper identification of traits such as color, texture, and shape of abnormal elements on plant leaves is crucial for the timely recognition of crop diseases.

Image segmentation is the process of selecting and grouping pixels with similar visual and numerical characteristics; that is, the separation of an image into regions. There are several criteria for selecting regions, segmentation can be performed in grayscale or color, and be bilevel or multilevel according to the number of the regions obtained from such segmentation [27].

For plant disease recognition, color features in the affected leaf regions are relevant and contribute to identifying shape and texture patterns that characterize each disease. In color segmentation, different models or color spaces are used that provide a standardized specification of color in a three-dimensional coordinate system and a subspace of the system, where each color is determined by a point. In this work, the RGB and HSV color spaces are used.

On the other hand, features to describe an object can be represented by boundaries or external properties, and by a structural representation or internal properties. It is desirable that the features are invariant to scale, rotation, and translation.

Textural features refer to the spatial distribution of gray hues defined by the uniformity, density, thickness, roughness, regularity, intensity, and directionality of discrete measures of hue and their spatial relationships. One approach of texture is based on the gray level co-occurrence matrix (GLCM), which is a frequency matrix with which a pixel with gray level (i) appears in a specific spatial relationship with another pixel of gray level (j). Concurrence matrices are second-order measures because they consider pairs of neighboring pixels, separated by a distance δ and at a given angle θ.

Color features are statistical measures of the color bands of a specific color space. Commonly used are mean, standard deviation, and entropy, among others. Regarding the shape features, the parameters to consider are the size and quantity of the regions of interest, circularity measures, perimeter, etc.

3 Methodology

AIn this section, the methodology for plant disease recognition is presented. First, we describe the proposed segmentation method. In the second phase, the feature extraction process and classification are presented.

3.1 Color segmentation

The first segmentation stage tries to visually identify the regions of interest (ROIs) on the plant leaf. Thus, four regions (R_i) are defined, where R_1 = plant leaf, R_2= shade, R_3= yellow-brown spots, and R_4= Hispa class spots. Figure 1 shows these regions.

Fig. 1. Regions of interest in the plant image.

The region R_4 is considered because the hispa class has greenish-whitish spots, unlike the rest of the classes that have spots ranging from light yellow to brown, as shown in Figure 2.

The visual identification of regions R_i is important for the segmentation of the whole leaf (without shadows), as well as for the segmentation of the regions representing the disease. Thus, the first step is to obtain thresholds for the segmentation. To this end, datasets are created with 20 cropped images obtained from each R_i. The images are obtained by manually cropping the ROI from the original RGB leaf image.

Fig. 2. Images of three classes of plants diseases. a). Brown spot, b) Leaf blast, c) Hispa

Considering that the H and S bands of the HSV color space represent the hue and color saturation information, in a second step, the RGB images are converted to the HSV color space. Using the H and S bands we create overlapping histograms of each dataset. The thresholds for segmenting each R_i region are obtained from the histogram information. Table 1 shows the results.

Table 2. Features extracted for classifiying.

Texture	Color	Shape
1.Uniformity 2.Entropy 3.Dissimilarity 4.Inverse difference 5.Correlation 6.Contrast 7. Inverse difference moment	1.Mean 2.Entropy 3.Kurtuosis 4.Skewness 5.Standard deviation	1.Area max 2.Num.Areas 3.Sum areas 4.Standard deviation of areas size

Of the thresholds obtained, it should be noted that the tone distribution for R_2 includes the entire range for H, while for S it represents the complement for the thresholds of the remaining regions, so these thresholds are omitted. In addition, the thresholds for R_1 and R_4 are similar, so we exclude the threshold for R_4. This is because the hispa class shades are light green. Consequently, the thresholds for leaf segmentation are $(0 \leq H \leq 0.5)$ & $(0.46 \leq S \leq 1)$, and for diseased region segmentation are $(0.17 \leq H \leq 0.5)$ & $(0.5 \leq S \leq 1)$.

Segmented images of the whole leaf and diseased regions are used for feature extraction at later stages for disease recognition. Figure 3 shows the scheme of proposed segmentation.

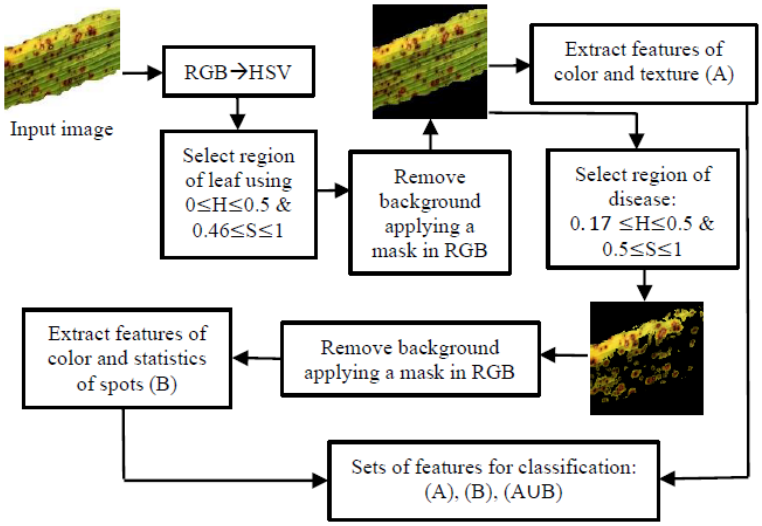

Fig. 3. Color segmentation proposed.

3.2 Feature extraction and Classification

The features used in this work are obtained independently for the segmented images of the whole leaf and the segmented images of the regions affected by the disease (spots). The features extracted for the leaf are color and texture (set A), while for the spots they are color and shape (set B). In the first case, these features are chosen from the results of previous work, which indicate that these features are suitable for leaf recognition [28], texture features are obtained from the gray level co-occurrence matrix (GLCM) using 8 gray levels, an interpixel distance of 1 and angles of 0, 45, 90, and 135 degrees.

Regarding the characteristics of the spots, color is suggested for its importance in visual identification, and shape due to the presence of patterns such as the number and size of affected regions. Table 2 shows the color, texture, and shape features used in this work.

Table 1. Thresholds for the interest regions R_i.

R_i	**H**	**S**
R_1 = plant leaf	0.00-0.50	0.46-1.00
R_2= shade	**0.00-1.00***	0.00-0.56
R_3= yellow-brown spots	0.17-0.50	0.50-1.00
R_4= Hispa class spots	0.00-0.50	0.45-1.00

From the extracted features, three sets are constructed and used independently for classification. The first set (A) consists of the color features for the H and S bands, respectively, furthermore the texture features

This set is obtained from the segmented image of the whole leaf. Thus, 17 features are obtained for the set A. The second set (B) is obtained from the segmented images of the leaf spots and is formed from the H and S bands color features, respectively, and shape features. Therefore, this set includes 14 features. A third set (AB) is formed including the features of sets A and B, totaling 31 features. Finally, the previously proposed sets are used for classifying, with three machine learning models: a multilayer perceptron neural network (MLP), a support vector machine (SVM) and random forest (RF).

4 Experiments and Results

This section describes the experiments performed and a brief discussion of the results obtained. A testbed for feature analysis using machine learning techniques is shown. Experiments were conducted on a computer with i9-7900X CPU 3.31Ghz, 64RAM, Windows 10 system, and Matlab 2018a.

4.1 Datasets

The dataset consists of images taken from two repositories of the Kaggle platform [10, 11]. The classes considered are brown spot, leaf blast, hispa with 560 images per class. The leaf blast and hispa class images are the result of a random selection of images from the repository [11], while the brown spot class set is composed from [10] and [11]. The images were resized to a size of 320 × 320 with the bicubic interpolation method since their original resolution is variable. Figure 2 shows examples of these images.

Training and test sets for each class use 80% and 20% of the images, respectively. This is 1344 training and 336 test images.

4.2 Color Segmentation

According to the segmentation method described in section 3.1, the results for each of the diseases addressed in this work are shown in Figures 4-6.

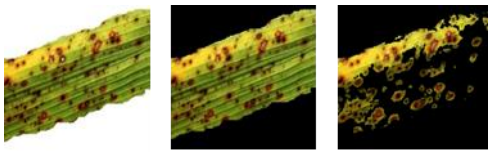

Fig.4 Segmented image of Brown spot class

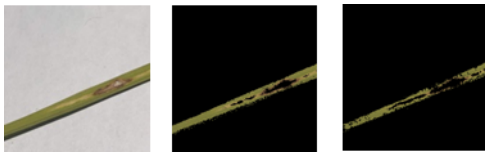

Fig.5 Segmented image of Leaf blast class

Fig.6 Segmented image of Hispa class

As can be seen, segmentation of the whole leaf is adequate for all classes, however, segmentation of disease regions (spots) in the Leaf blast and Hispa classes results in some regions being part of the leaf, or else lost as in the the Hispa class. The results for the Brown spot class are the best due to the prominence of the affected regions.

4.3 Classification

Experiments of classification are performed 15 times for a more objective evaluation Normalization used is z-score due to its simplicity and the possibility of comparing data sets coming from different distributions.

The classifiers used in all experiments are MLP, SVM, and RF. The network topology used is three hidden layers with the configuration [30,12,30] obtained by experimentation. The solver used is 'trainlm' and 60 epochs. The SVM is nonlinear and uses a Gaussian kernel, while the RF is initialized with 500 trees, and classification is by majority vote [3].

The experiments include the evaluation of sets A, B, and AB with the three classifiers. In addition, considering that set AB has 31 features, experiments are performed using the Relief algorithm [29] for feature selection. Tables 3 and 4 show the results.

Table 4. Test accuracy (%) for 15 executions of set AB using Relief algorithm and different quantity of features (the best)

Statistics	20 feats	23 feats	25 feats
Max (MLP)	73.81	75.30	74.40
Std (MLP)	1.57	2.44	2.19
Media (MLP)	71.45	71.39	70.58
Max (SVM)	70.54	71.13	71.73
Std (SVM)	0.63	0.91	0.73
Media (SVM)	69.23	70.08	70.56
Max (RF)	75.60	77.08	77.68
Std (RF)	0.83	1.19	1.18
Media (RF)	74.37	73.93	74.66

As seen from the results in Table 4, feature selection for the AB set has little relevance on the classification results with slight improvements using a set of 25 features: for SVM from 69.42% to 70.56%, and for RF from 74.44% to 74.66%. It is also observed that using a smaller number of features in most cases decreases the classification rate. In all cases, the features omitted

due to their low importance are dissimilarity, contrast, leaf correlation, and kurtuosis of the H and S bands, respectively of the spots. Moreover, in both experiments (with and without feature selection), the AB set provides better classification results.

It is worth mentioning that the purpose in this work is to identify the most appropriate features for the recognition of leaf diseases, and not to exceed the classification accuracy with respect to other works.

Table 3. Test accuracy (%) for 15 executions of sets A, B, AB

Statistics	MLP	SVM	RF
Max (A)	74.70	71.43	76.19
Std(A)	1.61	0.69	1.02
Media (A)	71.77	70.40	74.42
Max (B)	67.86	64.58	66.67
Std(B)	2.11	0.67	1.16
Media (B)	63.81	63.13	64.27
Max (AB)	74.11	70.24	75.89
Std (AB)	1.77	0.78	0.79
Media (AB)	71.47	69.42	74.44

Moreover, it is not the intention to establish comparisons with other learning approaches such as deep learning or generative neural networks [4,13,14].

Finally, an important purpose in this work is the use of expert knowledge for disease recognition in plant leaves, consequently, classical supervised learning techniques such as MLP, SVM and RF are addressed. In this sense, visual features that human experts look for in plant leaves to identify the disease are considered in a punctual way, for which the attention is towards the extraction of features in the regions of interest in the leaf and not in the whole image, as in the case of deep learning techniques based on neural networks. With respect to the latter, high classification rates are often due to features external to the object(s) of interest in the image, exacerbating the problem of explainability in the learning model by going against what a human expert identifies as relevant information in an image for recognition.

5 Conclusions

In this work, we address the problem of recognition of diseases in the leaves of rice crops classical using color segmentation and machine learning. Unlike other proposals, we use balanced sets of images of an appropriate size for the models used. We also propose a set of experiments to analyze the features for disease leaf recognition. From the results, it is clear the need to identify more precise characteristics for the classes of leaves those present similarities. Particularly, in the classes analyzed in this work, a common problem is the extraction of features of the disease regions, which has an impact on the classification accuracy achieved by the classification models.

Some areas of opportunity for future work include the use of robust segmentation methods for images acquired in uncontrolled environments, sharing the dataset used in this work with the scientific community, the automatic selection of features to improve the classification rate, and automatic classification models that provide some degree of explainability concerning the results obtained.

Acknowledgements

The authors would like to acknowledge the support provided by the Autonomous University of Tlaxcala, Mexico. The authors also express their gratitude to the Applied Computational Intelligence Network (RedICA).

References

1. Majji, V. Applalanaidu, & G. Kumaravelan. (2021). A review of machine learning approaches in plant leaf disease detection and classification. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 716–724.
2. Bhagat, M., & Kumar, D. (2002). A comprehensive survey on leaf disease identification & classification. *Multimedia Tools and Applications*, *81*, 33897–33925.
3. Breiman, L. (2001). Random forest. *Machine Learning*, *45*(1), 5–32.
4. Chug, A., Bhatia, A., Singh, A., & Singh, D. (2022). A novel framework for image-based plant disease detection using hybrid deep learning approach. *Soft Computing*, *27*.
5. Secretaría de Agricultura y Desarrollo Rural. (2023). Corn, beans, rice, and wheat, are Mexico's staple grains. Available at: https://www.gob.mx/agricultura/articulos/maiz-frijol-arroz-y-trigo-losgranos-basicos-de-mexico.
6. Dubey, A., Batra, D., Kumar, G., Kumar, S., & Singh, M. (2023). Comparative analysis of machine learning methods for plant disease identification. In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 573–578.
7. Organización de las Naciones Unidas para la Agricultura y la Alimentación. (2023). Available at: https://www.fao.org/home/es.
8. Gómez-Camperos, J. A., Jaramillo, H. Y., & Guerrero-Gómez, G. (2022). Digital image processing techniques for detection of pests and diseases in crops: a review. *Ingeniería y Competitividad, 24*.
9. Hassan, S. M., Amitab, K., Jasinski, M., Leonowicz, Z., Jasinska, E., Novak, T., & Maji, A. K. (2022). A survey on different plant diseases detection using machine learning techniques. *Electronics*, *11*(17).
10. Rice Leaf Diseases. (2023). Available at: https://www.kaggle.com/datasets/vbookshelf/rice-leafdiseases.
11. Riceleafs. (2023). Available at: https://www.kaggle.com/datasets/shayanriyaz/riceleafs.
12. Kaur, S., Pandey, S., & Goel, S. (2018). Plants disease identification and classification through leaf images: A survey. *Archives of Computational Methods in Engineering*, *26*.
13. Moussafir, M., Chaibi, H., Rachid, S., Chehri, A., Abdessamad, R., & Jeon, G. (2022). Design of efficient techniques for tomato leaf disease detection using genetic algorithm-based and deep neural networks. *Plant and Soil*, *479*, 1–16.
14. Militante, S. V., Gerardo, B. D., & Dionisio, N. V. (2019). Plant leaf detection and disease recognition using deep learning. In *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, 579–582.
15. Mugithe, P. K., Mudunuri, R. V., Rajasekar, B., & Karthikeyan, S. (2020). Image processing technique for automatic detection of plant diseases and alerting system in agricultural farms. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, 1603–1607.
16. OCDE. (2023). What is the future of food and farming. Available at: https://www.oecd.org/agriculture/understanding-the-global-food-system/what-is-the-futureof-food-and-farming/.
17. The World Bank. (2023). Agriculture and Food. Available at: https://www.worldbank.org/en/topic/agriculture/overview.
18. Wani, J. A., Sharma, S., Muzamil, M., Ahmed, S., Sharma, S., & Singh, S. (2022). Machine learning and deep learning based computational techniques in automatic agricultural diseases detection: Methodologies, applications, and challenges. *Archives of Computational Methods in Engineering*, *29*, 641–677.
19. Shrivastava, V. K., & Pradhan, M. K. (2021). Rice plant disease classification using color features: A machine learning paradigm. *J. Plant Pathol*, *103*, 17–26.
20. Ramesh, S., & Vydeki, D. (2020). Recognition and classification of paddy leaf diseases using Optimized Deep Neural network with Jaya algorithm, *Information Processing in Agriculture*, *7*(2), 249–260. https://doi.org/10.1016/j.inpa.2019.09.002.
21. Phadikar, S., Sil, J., & Das, A. K. (2013). Rice diseases classification using feature selection and rule generation techniques. *Comput. Electron. Agric*, *90*, 76–85.
22. Phadikar, S., & Sil, J. (2008). Rice disease identification using pattern recognition techniques. In *2008 11th International Conference on Computer and Information Technology*, 420–423. https://doi.org/10.1109/ICCITECHN.4803079.
23. Majid, K., Herdiyeni, Y., & Rauf, A. (2013). I-PEDIA: Mobile application for paddy disease identification using fuzzy entropy and probabilistic neural network. In *Proceedings of the 2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 403–406.
24. Pugoy, R. A. D., & Mariano, V. Y. (2011). Automated rice leaf disease detection using color image analysis. In *Proceedings of the Third International Conference on Digital Image Processing (ICDIP 2011)*, International Society for Optics and Photonics, Chengdu, China; Volume 8009, p. 80090F.
25. Yao, Q., Guan, Z., Zhou, Y., Tang, J., Hu, Y., & Yang, B. (2009). Application of support vector machine for detecting rice diseases using shape and color texture features. In *Proceedings of the International Conference on Engineering Computation*, 79–83.
26. Prajapati, H. B., Shah, J. P., & Dabhi, V. K. (2017). Detection and classification of rice plant diseases. *Intell. Decis. Technol.*, 357–373.
27. González, R., & Woods, R. (2017). *Digital Image Processing* (4th ed.). Pearson.
28. Ochoa Montiel, R., Fernández Hernández, J. P., & Montalvo Galicia, F. (2024). Handcraft learning for leaf disease recognition in rice crops. *Abstraction & Application Revista Electrónica de la Facultad de Matemáticas Universidad Autónoma de Yucatán*, *44*(1), 139–148. ISSN: 2007-2635.
29. Kononenko, I., Šimec, E., & Robnik-Šikonja, M. (1997). Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence*, *7*, 39–55. https://doi.org/10.1023/A:1008280620621.