Application of Unsupervised Learning in the Early Detection of Late Blight in Potato Crops Using Image Processing

Aplicación del Aprendizaje No Supervisado en la Detección Temprana del Tizón Tardío en Cultivos de Papa mediante Procesamiento de Imágenes

DOI: http://doi.org/10.17981/ingecuc.18.2.2022.07

Artículo de Investigación Científica. Fecha de Recepción: 13/09/2022. Fecha de Aceptación: 20/09/2022.

Juana-Valentina García-Ariza
Universidad Pedagógica y Tecnológica de Colombia. Sogamoso (Colombia)
juana.garcia01@uptc.edu.co

Marco-Javier Suarez-Barón
Universidad Pedagógica y Tecnológica de Colombia. Sogamoso (Colombia)
marco.suarez@uptc.edu.co

Edmundo-Arturo Junco-Orduz
Universidad Pedagógica y Tecnológica de Colombia. Sogamoso (Colombia)
edmundo.junco@uptc.edu.co

Juan-Sebastián González-Sanabria
Universidad Pedagógica y Tecnológica de Colombia. Tunja (Colombia)
juansebastian.gonzalez@uptc.edu.co

To cite this paper:
J. García-Ariza, M. Suarez-Barón, E. Junco-Orduz & González-Sanabria, "Application of Unsupervised Learning in the Early Detection of Late Blight in Potato Crops Using Image Processing", *INGE CUC*, vol. 18, no. 2, pp. 89–100. DOI: http://doi.org/10.17981/ingecuc.18.2.2022.07

Resumen

Introducción— La detección automática puede ser útil en la búsqueda de grandes campos de cultivo simplemente detectando la enfermedad con los síntomas que aparecen en la hoja.

Objetivo— Este artículo presenta la aplicación de técnicas de aprendizaje automático destinadas a detectar la enfermedad del tizón tardío utilizando métodos de aprendizaje no supervisados como K-Means y agrupamiento jerárquico.

Método— La metodología utilizada está compuesta por las siguientes fases— adquisición del dataset, procesamiento de la imagen, extracción de características, selección de características, implementación del modelo de aprendizaje, medición del rendimiento del algoritmo, finalmente se obtuvo una tasa de acierto del 68.24% siendo este el mejor resultado de los algoritmos de aprendizaje no supervisados implementados, usando 3 clusters para el agrupamiento.

Resultados— De acuerdo con los resultados obtenidos, se puede evaluar el desempeño del algoritmo K-Means, es decir, 202 aciertos y 116 errores.

Conclusiones— Los algoritmos de aprendizaje no supervisado son muy eficientes al momento de procesar una gran cantidad de datos, en este caso una gran cantidad de imágenes sin necesidad de etiquetas predefinidas, su uso para solucionar problemas locales como afectaciones de tizón tardío en cultivos de papa es novedoso.

Palabras clave— Aprendizaje automático; aprendizaje no supervisado; K-Means; agrupamiento jerárquico; tizón tardío

Abstract

Introduction— Automatic detection can be useful in the search of large crop fields by simply detecting the disease with the symptoms appearing on the leaf.

Objective— This paper presents the application of machine learning techniques aimed at detecting late blight disease using unsupervised learning methods such as K-Means and hierarchical clustering.

Method— The methodology used is composed by the following phases— acquisition of the dataset, image processing, feature extraction, feature selection, implementation of the learning model, performance measurement of the algorithm, finally a 68.24% hit rate was obtained being this the best result of the unsupervised learning algorithms implemented, using 3 clusters for clustering.

Results— According to the results obtained, the performance of the K-Means algorithm can be evaluated, i.e. 202 hits and 116 misses.

Conclusions— Unsupervised learning algorithms are very efficient when processing a large amount of data, in this case a large amount of images without the need for predefined labels, its use to solve local problems such as late blight affectations in potato crops are novel.

Keywords— Machine learning; unsupervised learning; K-Means; hierarchical clustering; late blight

I. Introduction

Potato is one of the most important crops in Colombia with an annual production of 2.8 million tons [1], one of the problems that most affect potato production is late blight, this is a disease caused by the fungus Phytophthora infestans [2], and losses of up to 15% of potato crops are attributed to it [3]. It should be clarified that this disease is not exclusive to potato and affects other crops such as maize, tomato, banana among others [4], this is a disease that can be treated in early stages [5], it is also easily identifiable by observing the characteristics of the leaves [6].

Normally an expert is needed to diagnose the disease in crops by observation of the leaves, this is an inaccurate method as there are other diseases that can affect potato leaves, causing an erroneous diagnosis [6]. Currently there are numerous studies that focus on the detection of this disease by signal processing and the use of artificial intelligence, more specifically supervised learning methods are used with many different architectures [7], [8]. The problem with these supervised methods is that they need a large number of images previously labelled as “healthy”, “diseased”, “onset of disease”, etc [6], which makes the training of artificial intelligence significantly time-consuming, and the training done for one potato variety will probably not be valid for another variety, so the time-consuming training process would need to be done again [8].

As far as unsupervised learning methods are concerned, it is evident that very little research exists for the detection of late blight disease [5], [9], [10], and even more so in potato cultivation. The advantage of using unsupervised learning methods is that it is not necessary to label the images for training the software, which makes the training of the algorithm faster, so that the algorithm can easily be trained for different populations and varieties of potato crop. This work shows the development of a software capable of detecting late blight disease in early (curable) stages, by using image processing techniques and the unsupervised learning methods K-Means and hierarchical clustering; in addition, a measurement and comparison of the efficiency of the software for the detection of the disease for the two previously mentioned methods is carried out.

II. Materials and Methods

The development of this research is designed in 5 phases comprising, image set acquisition, processing, feature extraction, clustering and finally model validation.

A. *Data Set*

For the acquisition of the set of images, the photographs were taken with a Samsung Galaxy A30 mobile device which has a 16MP camera with an aperture of f/1. 7 and Phase Detection Auto Focus (PDAF) technology; the samples were taken in different potato crops in two different areas, in the municipalities of Tuta and Aquitania, Boyacá (Colombia). As a first step, healthy leaves and leaves showing symptoms of late blight disease on potato leaves were identified, these leaves were collected and a photographic record was made on a white background; samples of leaves in different states (healthy, beginning of the disease and with advanced disease) were obtained, with a total of 320 images captured.

The whole set of images is stored in a folder in random order. In Fig. 2 you can see some of the photographs of the potato leaf in three states: healthy, initiating disease and advanced disease. The whole set of images is stored in a folder in random order. (Fig. 1).

Fig. 1. Images captured, from left to right, healthy, early disease and diseased.
Source: Authors.

Next, an exploratory analysis of the images taken of the leaves is presented. The following graph shows the classification of all the images used, in this case there are 154 diseased leaves, 26 at the beginning of the disease and 140 healthy leaves (Fig. 2a). It can also be seen that more than half of the samples correspond to *pastusa potato*, according to the grouping of the samples by potato variety of the crops used (Fig. 2b).

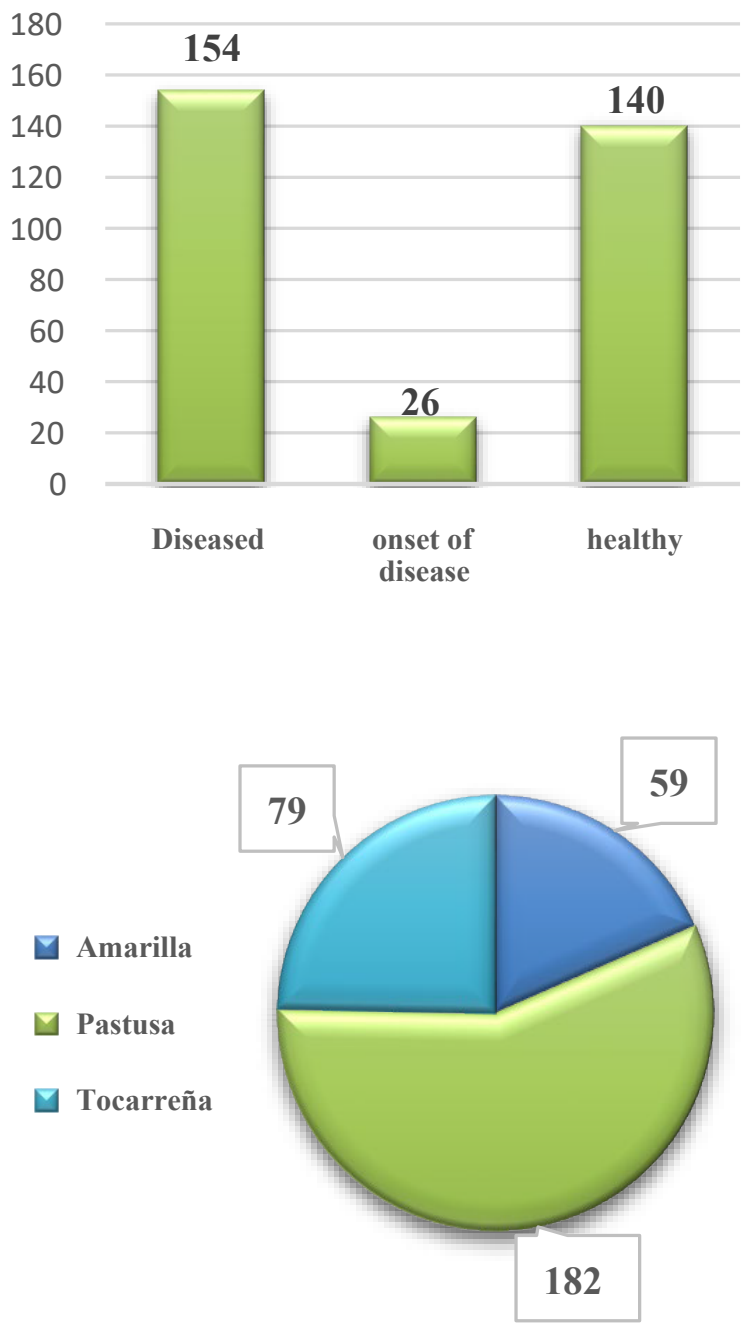

Fig. 2. Number of potato leaves according to (a) condition and (b) variety.
Source: Authors.

B. *Image Processing*

In this phase, color detection in the HSV color space is used, defining the minimum and maximum ranges to filter the image using a channel of a specific color, this is done with the OpenCv library. The image of the potato leaf is segmented, filtering the image and leaving only most of the pixels: green (healthy part) and brown (the spots on the leaf), in order to generate a mask that eliminates the background of the image. It is necessary to know the range of colors to filter from the image, to find the correct lower range of the green or brown color in the H channel, first an approximate initial value of H is assigned, corresponding to the green or brown colour (the whole procedure must be carried out for both colours) and the flow of the algorithm is followed until reaching the S channel, where the same procedure is carried out and the minimum value of the V channel is found, then the procedure is started again for the missing colour (green or brown) (Fig. 3a). The algorithm for finding the maximum values of both green and brown in the H, S and V channels (Fig. 3b).

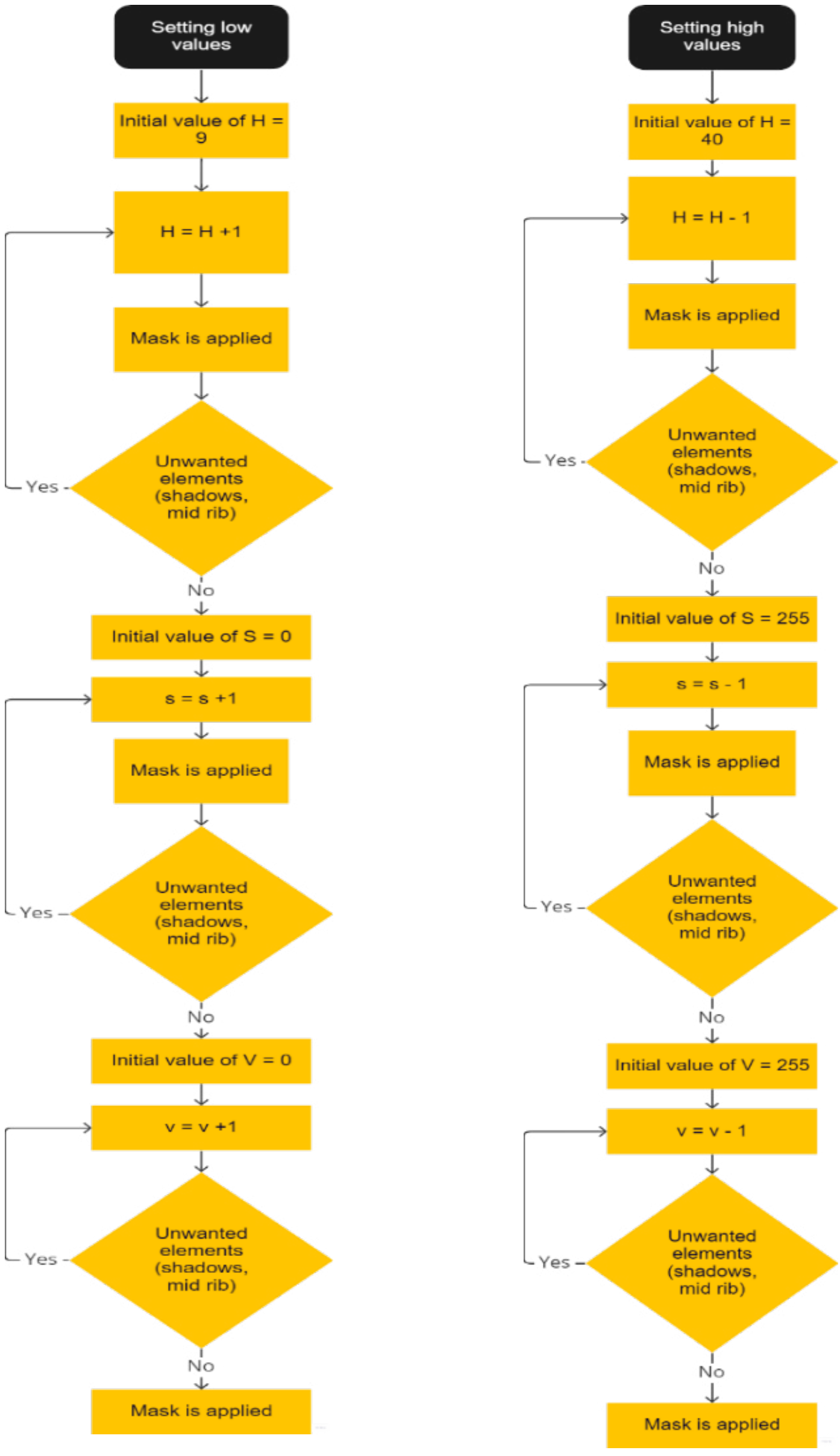

Fig. 3. Algorithms for finding (a) minimum and (b) maximum ranges.
Source: Authors.

The following diagram summarizes the segmentation procedure by means of color detection in the HSV colour space (Fig. 4). This algorithm was chosen to remove the background from the images in order to extract the necessary features to feed the unsupervised learning method; initially the image is resized, then it is converted from the BGR color space to the HSV space (this is because the HSV space is more intuitive and easier to use), then according to the previously established colour ranges the green (healthy leaves) and light brown and dark brown (diseased leaves) masks are applied, the three results obtained are printed on the screen, finally these three masks are merged into one image and printed, in this image the original image should be seen but without the background.

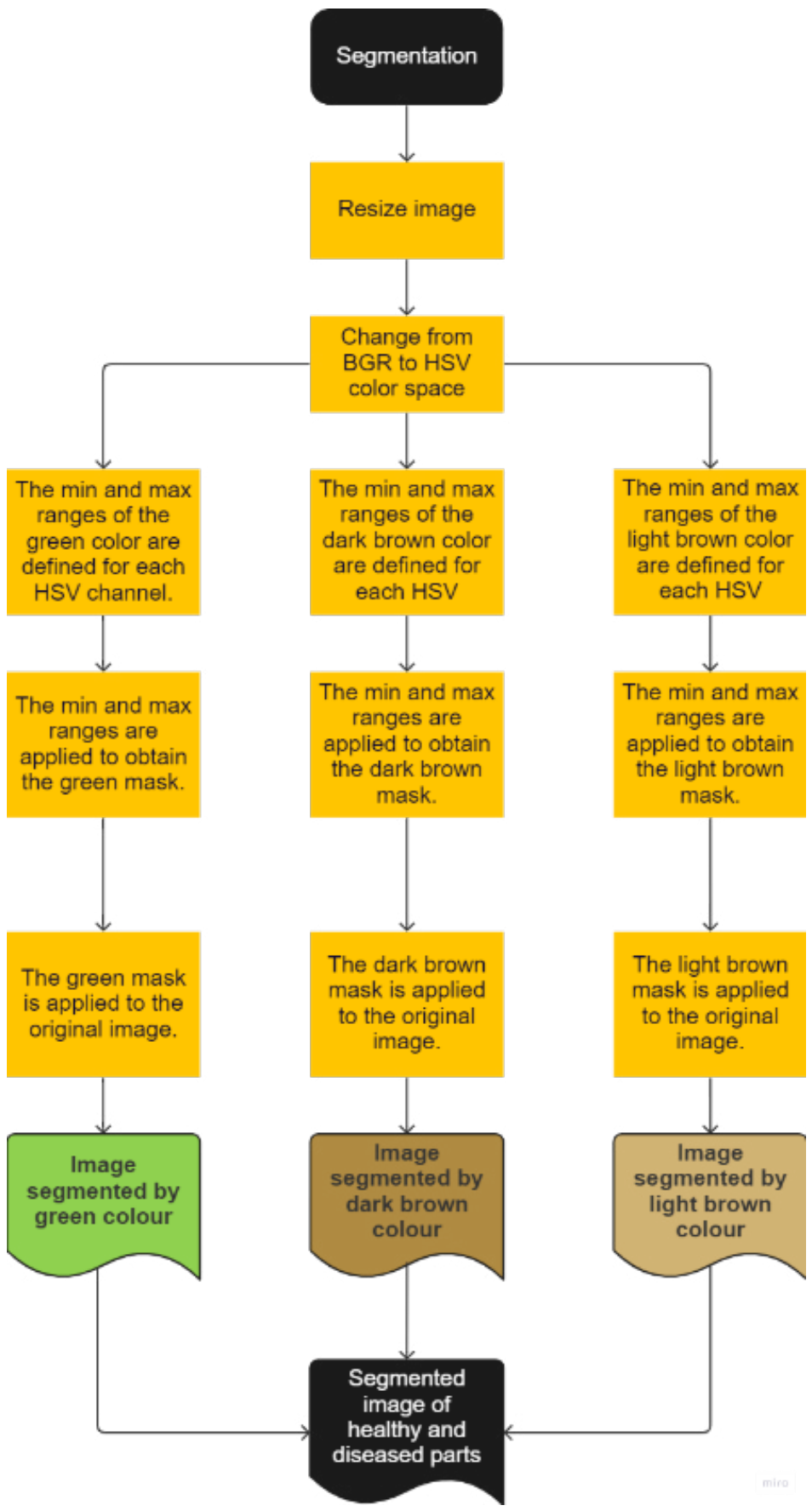

Fig. 4. Segmentation flowchart.
Source: Authors.

The result of applying this segmentation algorithm can be seen in the images of a diseased leaf (Fig. 5a), leaf initiating disease (Fig. 5b) and healthy leaf (Fig. 5c).

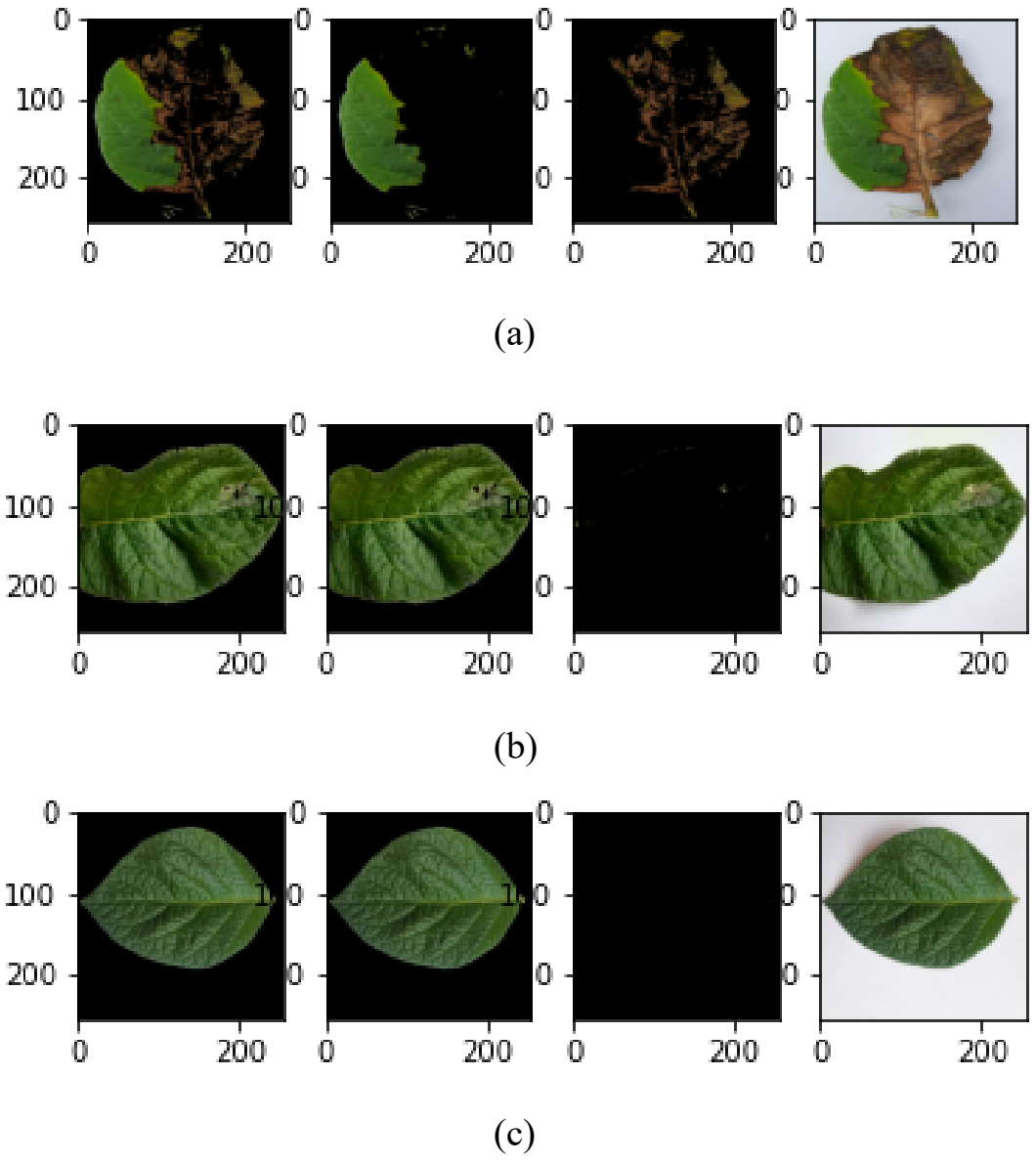

Fig. 5. Segmentation into (a) diseased leaf, (b) leaf initiating disease and (c) healthy leaf.
Source: Author.

C. *Feature extraction*

In this phase two types of features are used to represent the digital images, the first are the color features by calculating statistical variables such as the mean, standard deviation, variance and the data range (the difference of the maximum and minimum value), these are applied to each channel of the BGR and HSV color space, giving a dimension of 24 features for each sample.

The following features are texture features using GLCM technique, the following features were implemented: contrast, dissimilarity, homogeneity, energy, correlation and ASM. The level-gray co-occurrence matrix is created, for which the parameters such as image, displacement every 1 pixel and angle direction 0, 90, 45 and 135 degrees are defined. Each of the features are calculated by receiving the generated GLCM function, a dimension of 24 features is obtained for each sample.

The last features calculated are the total number of pixels in the stain and the ratio between healthy and diseased pixels. To determine the total number of pixels, a sum of the count of pixel values equal to 1 in each of the brown masks generated in the segmentation was performed. If the total number of sick pixels equals zero, a value of 0.1 is assigned, since a division by 0 can be generated when performing the pixel ratio. To generate the pixel ratio (pse ratio) is used (1), since, if the leaf image is captured from very far away or very close this pixel ratio should not be affected, unlike the number of green pixels and the number of brown pixels which are directly affected by the distance at which the leaf image is captured.

$$pse\ ratio \quad \frac{Total\ green\ pixels}{Total\ brown\ pixels} \tag{1}$$

D. *Clustering*

1) K-Means

The K-Means algorithm divides the data and assigns n samples to one of the k clusters defined by the centroids, for which the first step is to search for or define a value of k, for this case a value of k equal to 3 is chosen, it is assumed that these clusters will be recognized by the algorithm as the states of the leaf, healthy, with advanced disease and starting the disease, although this must be analyzed in more depth in the results obtained. Before running the training, the number of maximum iterations to be performed to find the centers of these groups is determined, when the cluster assignment for each sample does not change. The sklearn.clustering library is used to adjust and predict the closest cluster to each sample. Flowchart for the K-Means algorithm (Fig. 6).

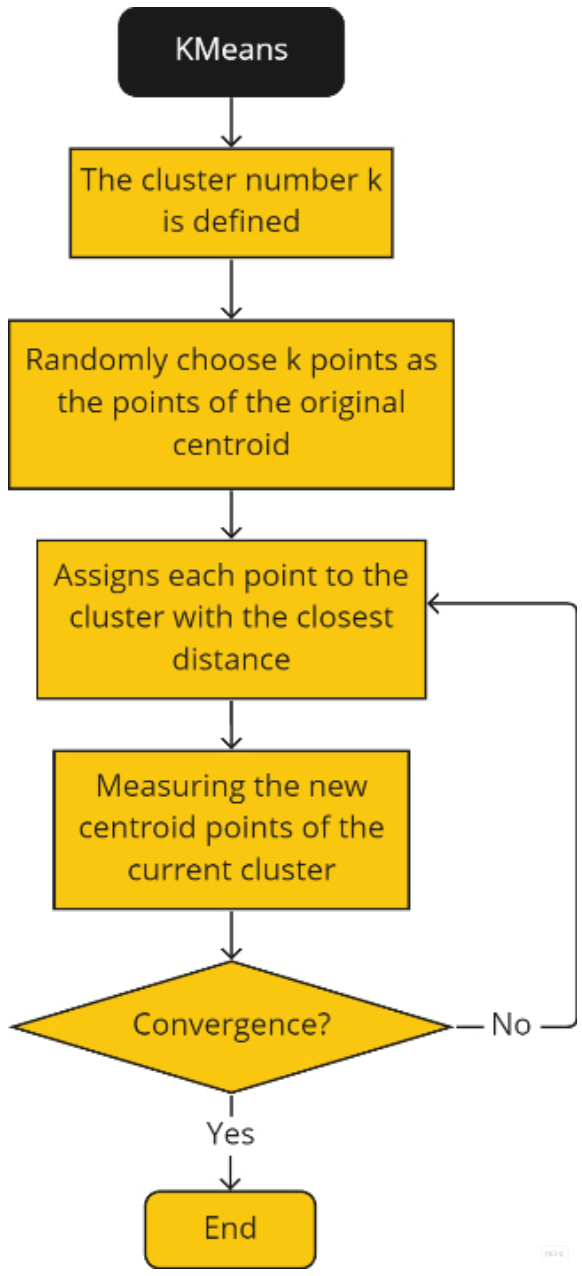

Fig. 6. Diagram for K-Means algorithm.
Source: Authors.

2) Hierarchical clustering

To apply the hierarchical agglomerative clustering algorithm, the first step is to define the number of clusters, the same number of clusters as those used in K-Means is used, that is 3, this is to be able to directly compare the results obtained, then the linkage parameter is defined, in this case the test is done for: ward and average, it is also necessary to define the affinity or distance metric to calculate the clustering in all cases the Euclidean (Euclidean) is used. To run the algorithm the agglomerative Clustering function from the sklearn.cluster library is used, to assign the clustering for each sample, the general hierarchical clustering algorithm is used (Fig. 7).

It is assumed that for each sample a group is created $G_i = \{x_i\}$, $i = 1 \ldots N$, $N > 1$.

$$G = \{G_1 \ G_{2, \ldots, GN}\}$$

Where the whole set of groups is represented by G.

In the set G, the closest groups G_k y G_j are searched. Then, these two groups are joined together creating the group G'_k. The size d of the group G'_k is calculated as the maximum distance between any two data in the group.

If $d >$ maxd then the group G'_k is removed. The grouping stops and the set G will be equal to the groups created.

Otherwise G_k is replaced by G'_k. The group G_j is removed from the set of groups G, the process is repeated to find the nearest groups. The grouping is stopped and the set G will form the created groups [11].

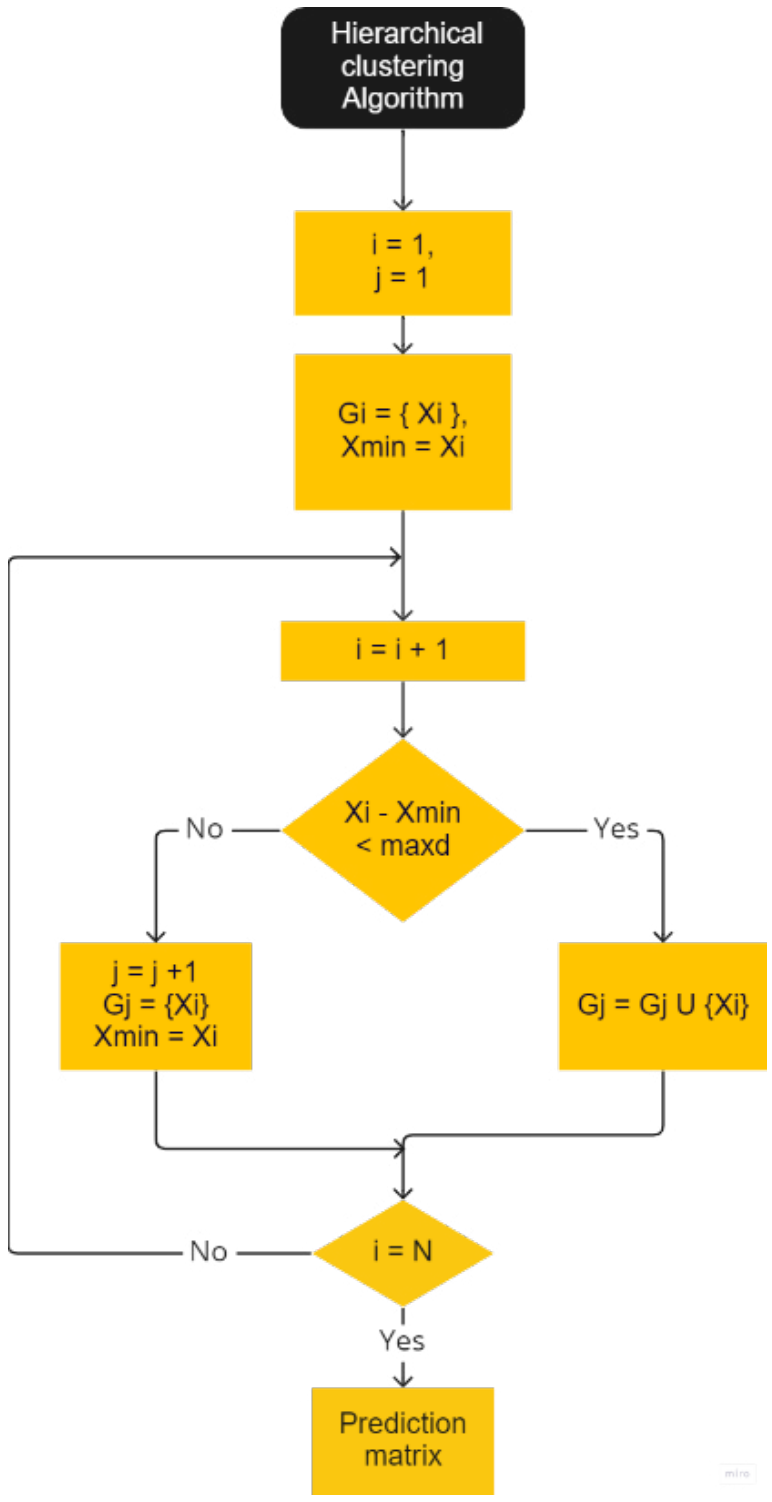

Fig. 7. Hierarchical clustering algorithm.
Source: Authors.

The dendrogram (Fig. 8) of the hierarchical clustering is implemented, which provides a graphical representation of the clustering performed by the algorithm, where each sample initially forms a group and according to levels of similarity, agglomerations are formed, generating a global tree. The optimum number of clusters to be generated is identified as 2 to 4 clusters.

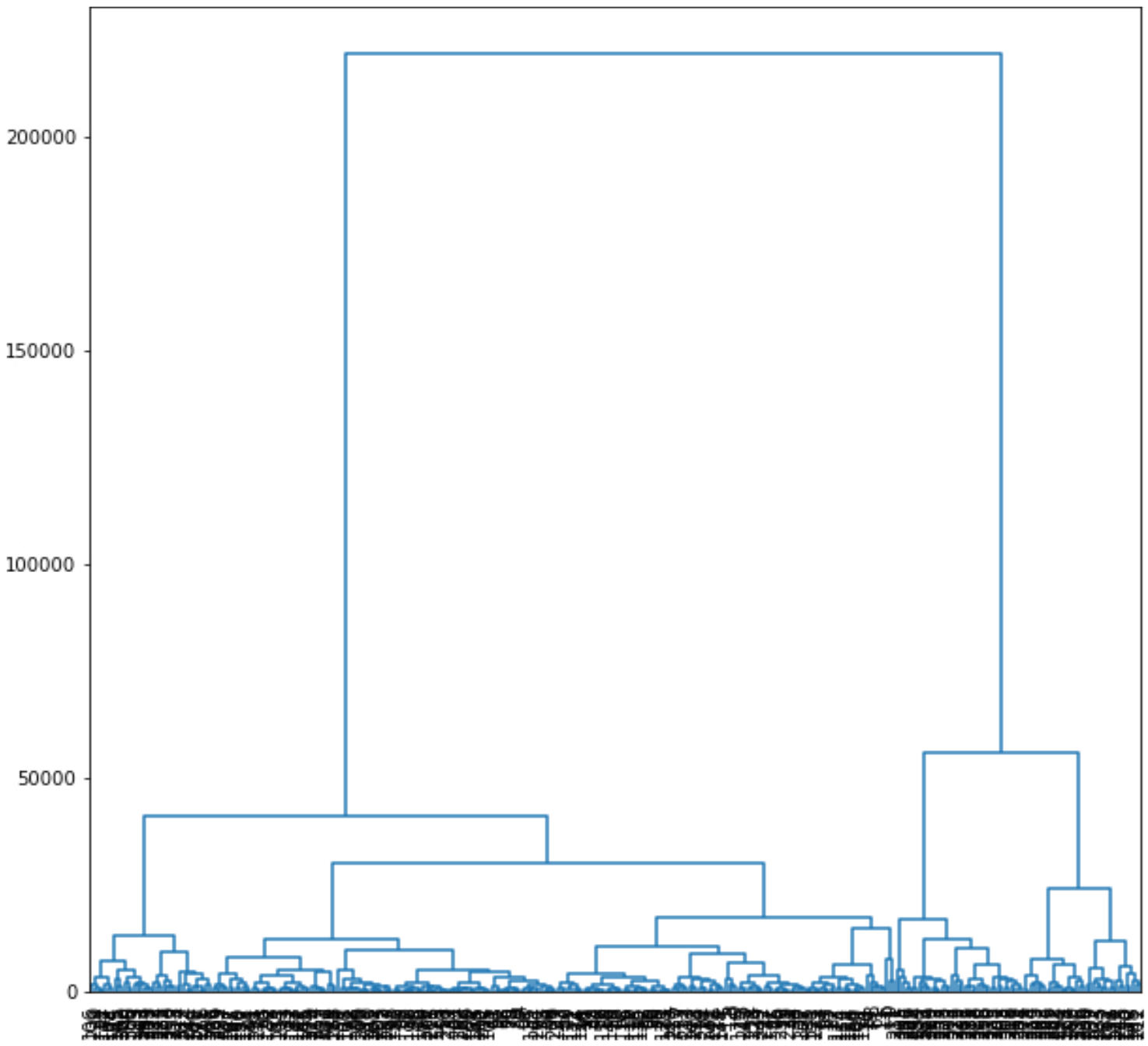

Fig. 8. Dendogram.
Source: Authors.

The following image shows the proposed methodology (Fig. 9).

Fig. 9. Proposed methodology.
Source: Authors.

III. Results and discussion

According to the results obtained, the performance of the K-Means algorithm (Fig. 10a) can be evaluated, i.e. 202 hits and 116 misses. In the case of the Ward hierarchical algorithm (Fig. 10b), 202 hits and 116 misses are obtained, i.e. the same number of hits as the K-Means algorithm. This makes sense since the K-Means algorithm and the Ward hierarchical clustering try to minimize the same Euclidean cost function, so they are quite similar algorithms.

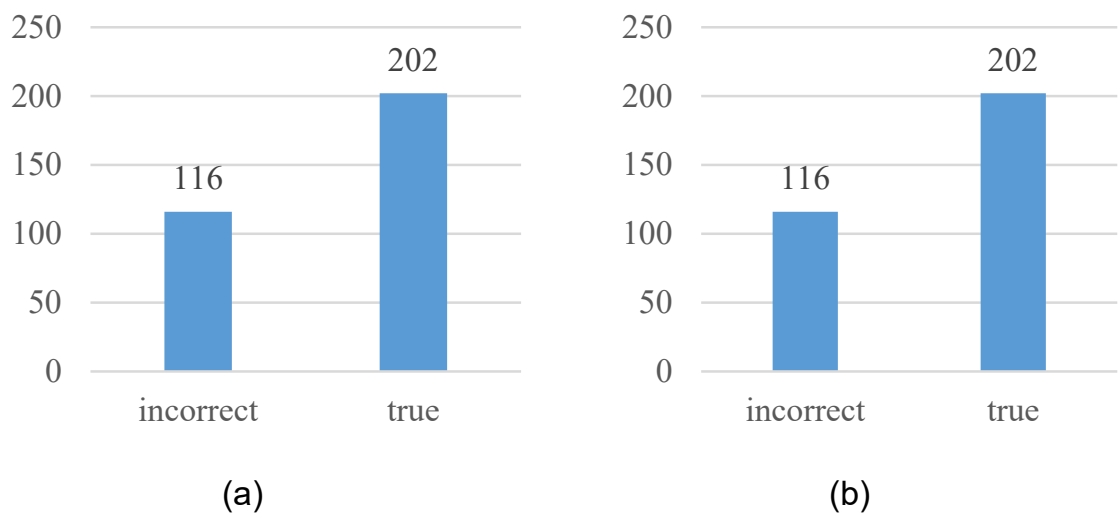

Fig. 10. Consolidated results of Ward hierarchical grouping.
Source: Authors.

Now for the Average hierarchical algorithm we obtain the results shown in the following figure, 217 hits and 101 misses (Fig. 11).

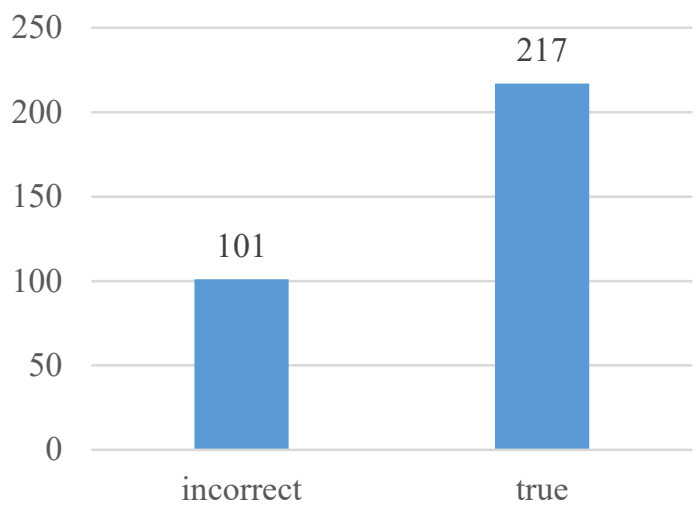

Fig. 11. Consolidated results hierarchical clustering average.
Source: Authors.

The figures obtained in Fig. 10 and Fig. 11 in percentages are shown in Table 1, it can be seen that the algorithm with the highest percentage of successes is the average hierarchical clustering with 68.23%, and with the same percentage of successes are the K-Means and Ward hierarchical clustering algorithms with 63.52%. Reviewing the state of the art, similar results are found, such as in the Pakistani work "Automatic detection of plant diseases; utilizing an unsupervised cascaded design" [12], the authors identify the late blight disease but in a banana crop with a hit percentage of 66.7% when using RGB and images with good illumination, while if the illumination is low they obtain a 50.4% success rate, it can be inferred that the application of the methods was correct; however, the photographs are a determining factor for a higher success rate.

Table 1.
Percentage of hits and misses of unsupervised training algorithms.

Algorithm	True (%)	Incorrect (%)
K-Means.	63.52	36.48
Hierarchical grouping average.	68.24	31.76
Hierarchical grouping ward.	63.52	36.48

Spurce: Authors.

A. *Dashboard design*

For the actual implementation in the field, prototypes of a mobile application are designed to allow the visualization of statistics, analysis of results and data behavior, which allows the administrator to obtain relevant information from the images and processed data. As shown in the following figures, different types of diagrams are used, such as bar diagrams to analyze the number of healthy and infected leaves processed per month and the number of healthy and infected leaves per phenological stage (Fig. 12a), circular diagrams to visualize the percentage of potato varieties and species processed (Fig. 12b) and finally prototypes for the different environmental conditions in which the crops are grown, such as temperature, humidity and precipitation (Fig. 12c).

Fig. 12. Prototypes design.
Source: Authors.

IV. Conclusions

Unsupervised learning algorithms are very efficient when processing a large amount of data, in this case a large amount of images without the need for predefined labels, its use to solve local problems such as late blight affectations in potato crops are novel, this work was developed from a set of data obtained locally in semi-controlled conditions to train three unsupervised learning algorithms (K-Means, Ward hierarchical clustering and average hierarchical clustering), Ward hierarchical clustering and average hierarchical clustering), it was quite difficult to segment the image and to correctly remove the background without losing important information for detection, finally it was achieved by means of a segmentation by means of color detection in the HSV space customized for this case.

The results obtained in terms of algorithm hits correspond to those found in the state of the art, in this case the best was 68.24%. It can be deduced that this could be improved if the images were captured with the leaf “alive”, i.e. not cut from the plant, in addition to the fact that all images should be captured with good lighting levels. Future work includes create and work with a dataset at least twice the size and created from “live” leaves, i.e. take the photograph with the image still attached to the plant; optimize the segmentation model using the HSV color space, implementing color filtering with a greater number of channels to make it more selective to late blight disease; implement the unsupervised learning algorithms in a mobile application that uses the dashboard designed and the mobile phone camera to capture the image and process it on site, and thus evaluate the performance of the algorithms designed in the field.

References

[1] Minagricultura, *Estrategia de ordenamiento de la producción cadena productiva de la papa y su industria*. BOG, CO: Minagricultura, 2019. Recuperado de https://sioc.minagricultura.gov.co/Papa/Normatividad/Plan%20de%20Ordenamiento%20papa%202019-2023.pdf

[2] C. Ortiz, "Desarrollo de una herramienta computacional basada en redes neuronales para el diagnóstico del tizón tardío en cultivos de papa", *Proyecto de grado*, Fac Ing Mec Electron Biomed, UAN, BOG, CO, 2021. Disponible en http://repositorio.uan.edu.co/handle/123456789/5156

[3] D. Rodríguez, M. Rico, L. Rodríguez y C. Ñústez, "Efecto de diferentes niveles y épocas de defoliación sobre el rendimiento de la papa (Solanum tuberosum cv. Parda Pastusa)," *Rev Fac Nal Agr MED*, vol. 63, no. 2, pp. 5521–5531, Sept. 2009. Disponibl en https://repositorio.unal.edu.co/handle/unal/37086

[4] A.-K. Mahlein, E.-C. Oerke, U. Steiner & H.-W. Dehne, "Recent advances in sensing plant diseases for precision crop protection," *Eur J Plant Pathol*, vol. 133, no. 1, pp. 197–209, Mar. 2012. https://doi.org/10.1007/s10658-011-9878-z

[5] S. Maity, S. Sarkar, A. Tapadar, A. Dutta, S. Biswas, S. Nayek & P. Saha, "Fault Area Detection in Leaf Diseases Using K-Means Clustering," presented *2nd International Conference on Trends in Electronics and Informatics*, ICOEI, TIRUN, IN, 11-12 May. 2018. https://doi.org/10.1109/ICOEI.2018.8553913

[6] J. Johnson, G. Sharma, S. Srinivasan, S. Masakapalli, S. Sharma, J. Sharma & V. Dua, "Enhanced field-based detection of potato blight in complex backgrounds using deep learning," *Plant Phenomics*, pp. 1–13, May. 2021. https://doi.org/10.34133/2021/9835724

[7] P. Sharma, Singh, B. & R. Singh, "Prediction of Potato Late Blight Disease Based Upon Weather Parameters Using Artificial Neural Network Approach," presented *9th International Conference on Computing, Communication and Networking Technologies*, ICCCNT, BLR, IND, 10-12 July 2018. https://doi.org/10.1109/ICCCNT.2018.8494024

[8] R. Hasan, S. Yusuf & L. Alzubaidi, "Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion," *Plants*, vol. 9, no. 10, pp. 1–25, Oct. 2020. https://doi.org/10.3390/plants9101302

[9] L. Li, S. Zhang & B. Wang, "Plant Disease Detection and Classification by Deep Learning - A Review," *IEEE Access*, vol. 9, pp. 56683–56698, Apr. 2021. https://doi.org/10.1109/ACCESS.2021.3069646

[10] H. Pardede, E. Suryawati, R. Sustika & V. Zilvan, "Unsupervised Convolutional Autoencoder-Based Feature Learning for Automatic Detection of Plant Diseases," presented *2018 International Conference on Computer, Control, Informatics and its Applications*, IC3INA, TANG, ID, 1-2 Nov. 2018. https://doi.org/10.1109/IC3INA.2018.8629518

[11] B. Małysiak-Mrozek, D. Mrozek & S. Kozielski, "Data Grouping Process in Extended SQL Language Containing Fuzzy Elements," in K.A. Cyran, S. Kozielski, J. F. Peters, U. Stańczyk & A. Wakulicz-Deja, *Man-Machine Interactions*, vol. 59, BE, DE, Springer, 2009, pp. 247–256. https://doi.org/10.1007/978-3-642-00563-3_25

[12] Z. Khan, T. Akram, S. Naqvi, S. Haider, M. Kamran & N. Muhammad, "Automatic detection of plant diseases; utilizing an unsupervised cascaded design," presented *15th International Bhurban Conference on Applied Sciences and Technology*, IBCAST, ISB, PK, 9-13 Jan. 2018. https://doi.org/10.1109/IBCAST.2018.8312246

Juana-Valentina García-Ariza. Universidad Pedagógica y Tecnológica de Colombia. Sogamoso (Colombia)

Marco-Javier Suarez-Barón. Universidad Pedagógica y Tecnológica de Colombia (Sogamoso, Colombia). https://orcid.org/0000-0003-1656-4452

Edmundo-Arturo Junco-Orduz. Universidad Pedagógica y Tecnológica de Colombia (Sogamoso, Colombia). https://orcid.org/0000-0002-9559-2146

Juan-Sebastián González-Sanabria. Universidad Pedagógica y Tecnológica de Colombia (Tunja, Colombia). https://orcid.org/0000-0002-1024-6077