An algorithm for characterizing skin moles using image processing and machine learning

Zaid Sanchez[1], Alicia Alva[2], Mirko Zimic[3], Christian del Carpio[4]
[1,4]Laboratorio de Investigación en Inteligencia Artificial, Robótica y Procesamiento de Imágenes,
Universidad Nacional de Ingeniería, Perú
[2,3]Laboratorio de Bioinformática y Biología Molecular, Universidad Peruana Cayetano Heredia, Perú

Article Info

Article history:

Received Jul 20, 2020
Revised Dec 12, 2020
Accepted Dec 22, 2020

Keywords:

Characterizing skin moles
Digital image processing
Melanoma
Random forest
Support vector machines

ABSTRACT

Melanoma, the most serious type of skin cancer, forms in cells (melanocytes) that produce melanin, the pigment that gives color to the skin. There are low-income regions that lack specialized dermatologists, causing skin cancer to be diagnosed in advanced stages. In Peru, in high Andean communities with low resources, the problem is aggravated by the high incidence of ultraviolet radiation and lack of medical resources to make the diagnosis. Normally, mole images are obtained from dermatoscopes. The present work seeks to use mole images obtained from smartphones to make the classification of them as suspected or not suspected of being melanoma, by means of a feature extraction algorithm. The first step is to make color and lighting corrections. After this, the image is segmented using the K-Means algorithm, and we obtain the areas of the mole and skin. With the segmented mole we proceed to extract the main visual characteristics and then use classification algorithms such as support vector machine (SVM), random forest and naïve bayes, which obtained an accuracy of 0.9473, 0.7368 and 0.6842, respectively. These results show that it is possible to use images obtained from smartphones to develop a classification algorithm with 94.73% accuracy to detect melanoma in skin moles.

Corresponding Author:

Christian del Carpio
Laboratorio de Investigación en Inteligencia Artificial, Robótica y Procesamiento de imágenes
Universidad Nacional de Ingeniería
Lima 15102, Perú
Email: cdelcarpiod@uni.edu.pe

1. INTRODUCTION

Skin diseases are the fourth leading cause of the global disease burden, affecting 30-70% of individuals and prevailing in all geographical regions and age groups [1]. The most common form of cancer in the United States is skin cancer, with about 5 million cases occurring annually [2-4]. Melanoma is the most dangerous type of skin cancer, causing more than 9.000 deaths per year [2, 3]. Although most melanomas are first discovered by patients [5], the diagnostic accuracy of unaided expert visual inspection is only 60% [6].

Melanoma is a cancer of the melanocytes and can develop as a new mole, or as part of a pre-existing mole [7]. According to the American Cancer Society, melanoma is less common than some other types of skin cancer, although it is more likely to grow and spread [8]. However, if detected in its early stage it can be treated effectively. For this reason, it is important to recognize the warning signs to watch for. ABCDE screening, according to Rigel [9] for early detection of this type of cancer, is a simple skin inspection method that allows the first signs to be determined so that one knows when to go to the dermatologist. This method, according to the American Academy of Dermatology, consists of checking the mole, with A being for asymmetry, B for the observation of borders, C for color, D for diameter and finally E for the evolution of the mole over time.

In this context, new algorithms have been developed in artificial intelligence for the classification of images [10], which have been applied to identify different diseases in the skin. These techniques are applied to dermatoscopic images [11, 12], which have high resolution and detail. L. Yu *et al.* in [13], developed a two-part algorithm; the first part is segmentation of the dermoscopic image and the second is classification. The first part uses a fully convolutional neural network (FCRN) for precise segmentation of skin lesions, and further enhances its capabilities by incorporating a multi-scale contextual information integration scheme and the second part uses a classification network. They processed 900 training and 350 test images, and obtained a 0.799 accuracy and a 0.844 segmentation Jaccard index.

Nasr-Esfahani *et al.* in [14], expose a classification algorithm using a data pre-processing, enhanced images are fed into a pre-trained convolutional neural network (CNN), which is a member of deep learning models. The CNN classifier, which is trained by a large number of samples, distinguishes between melanoma and benign cases. For 6120 images they obtain a sensitivity of 0.82, a specificity of 0.87 and an accuracy of 0.81.

In addition to the deep learning techniques, it is also possible to extract useful features from the images and use machine learning algorithms to perform the classification, in Mustafa *et al.* [15] it is based on the ABCDE features of the moles to perform the characterization of the images and in the end it uses an SVM classifier for the identification of the melanoma obtaining an accuracy of 86.67%.

In Murugan e*t al.* [16], they compare classification algorithms for the detection of skin cancer. The algorithms used are SVM, random forest and kNN classifier. For feature extraction they are also based on the ABCD rule, obtaining an accuracy of 89.43%, 76.87% and 69.54% for SVM, random forest and kNN classifiers respectively.

Arasi [17] proposes to use naive bayesian and decision tree techniques for the classification of images obtained from a dermatoscope, and to identify melanoma. This technique is also based on the use of the ABCD rule. Using only the texture characteristics in the analysis of the images, 84.5% accuracy is obtained when using the naive bayes classifier and 77.4% when using decision tree.

In all the works cited above, images obtained from a dermatoscope have been used to make the identification of melanoma in the skin. In this work we propose the possibility of using other sources of images, in particular we will use images obtained from smartphones, which have been captured by nurses in health centers in Peru. With this dataset we propose to extract characteristics of the mole based on the ABCD rule and use machine learning techniques such as SVM, random forest and naive bayes to make the identification of the melanoma. Machine learning was chosen instead of deep learning because in the articles cited it is concluded that its accuracy is similar to the deep learning methods and also has less computational load when processing.

2. RESEARCH METHOD

In this article we propose an algorithm to obtain the characteristics according to ABCD of a skin mole, in order to determine whether this mole is suspected of being a melanoma or not. As can be seen, the evolution of the mole over time is not taken into account. The block diagram of the proposed algorithm is shown in Figure 1.

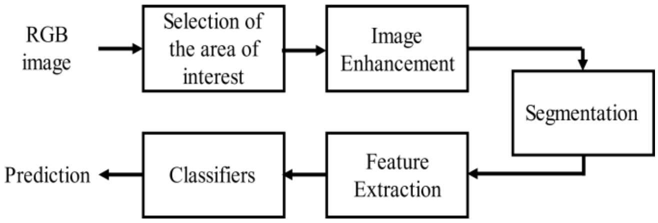


Figure 1. Block diagram of the proposed algorithm

2.1. Acquisition of the images

The images of the moles were captured with smartphones that were available at the health centers, the models were: Samsung S7 (resolution 12 MP), Samsung S6 (resolution 16 MP), ZTE Blade (resolution 15.9 MP) and LG K10 (resolution 13 MP). In each image a marker is added that has a size of 25x15 mm as shown in

Figure 2. All the images were acquired during the development of the MELap project that had a duration of 24 months during the years 2017 and 2018, carried out by the Universidad Peruana Cayetano Heredia. These images were obtained from 15 centers in which dermatological consultations are performed, located in Huancavelica, Arequipa, Tacna, Callao, Lima [18].

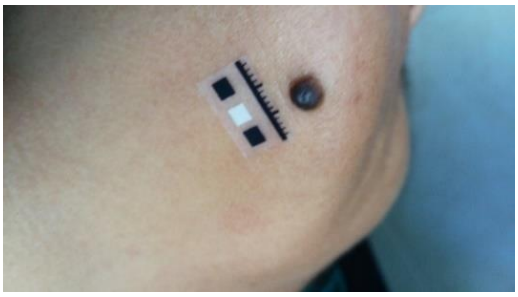


Figure 2. Initial image of the mole with marker

2.2. Selection of the area of interest

Because the entire image does not need to be processed, the area of interest in the image obtained initially, including the marker, is selected and then a perspective correction is performed [19]. A bilinear transformation is applied to the image to make the correction. Let A be the original color image and select 4 coordinates $(x_{1a}, y_{1a}), (x_{2a}, y_{2a}), (x_{3a}, y_{3a})$, (x_{4a}, y_{4a}), as shown in Figure 3(a).

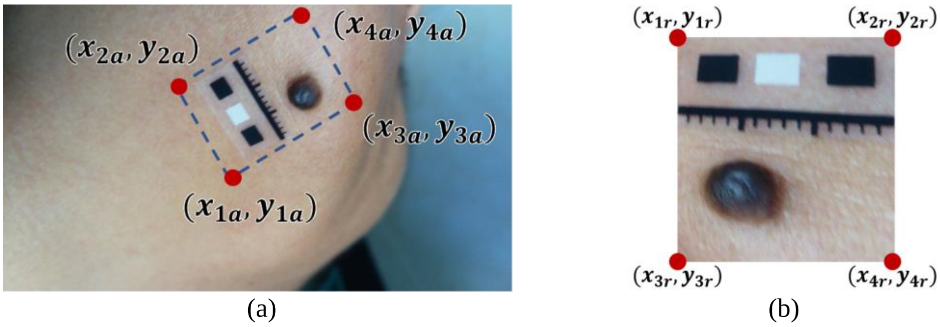


Figure 3. Selection of the area: (a) Selection of the 4 points on the picture, (b) Image after transformation size 300x300 pixels

To the selected coordinates, the change of perspective is applied and 4 new coordinates are obtained $(x_{1r}, y_{1r}), (x_{2r}, y_{2r}), (x_{3r}, y_{3r})$, (x_{4r}, y_{4r}) in the resulting picture R, as shown in Figure 3(b). In order to perform the perspective correction [19], a transformation matrix of 4x2 is defined, which transforms the coordinates of the original image and obtains the new desired coordinates as defined in (1).

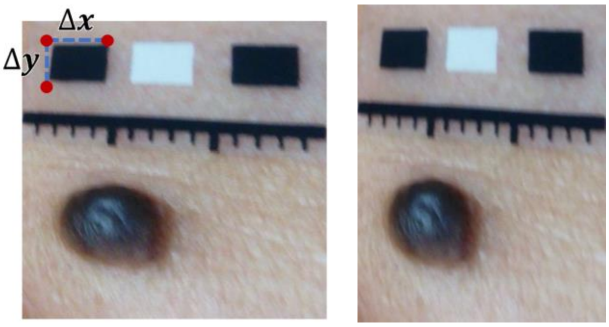
$$\begin{bmatrix} x_{ir} \\ y_{ir} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \end{bmatrix} \cdot \begin{bmatrix} x_{ia} \\ y_{ia} \\ x_{ia} \cdot y_{ia} \\ 1 \end{bmatrix} \tag{1}$$

where $i = 1,2,3,4$, indicates each of the four coordinates.

The transformation matrix has 8 unknowns to solve that would be $c_{11}, c_{12}, c_{13}, c_{14}, c_{21}, c_{22}, c_{23}, c_{24}$ and each pair of points produces 2 equations, so selecting 4 points on the image is enough to solve the equation. To the obtained image R that has a dimension of 300x300 pixels, another correction of perspective is made having as reference the squares that have the marker that was added to the image. It is known that each square measures $3x3$ mm, Figure 4(a) shows the image with that marker. The equivalence made is 1 mm equal to 20 pixels and using this correspondence the size of the final image is calculated using (2) and (3), obtaining the image R_A that has a dimension of $Tx \times Ty$ pixels, shown in Figure 4(b).

$$Tx = \frac{20\cdot 3\cdot 300}{\Delta x} \quad (2)$$

$$Ty = \frac{20\cdot 3\cdot 300}{\Delta y} \quad (3)$$



(a) (b)

Figure 4. Transformed with the equivalence of 1mm equal to 20 pixels; (a) R image, (b) R_A image of size $Tx \times Ty$

2.3. Image enhancement

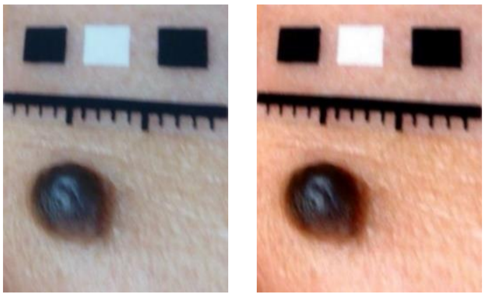
At this stage, an image intensity level correction is performed [20], because when the image is obtained, it may have different levels of illumination, which is not uniform. The marker in the image is used for this correction, as the black square will represent intensity level 0 and the white square in the marker will represent intensity level 255.

$$I = \frac{(R_A - Intensity_{min})\cdot 255}{Intensity_{max}} \quad (4)$$

where, $Intensity_{min}$ and $Intensity_{max}$, are the intensity of the black marker of the image and the intensity of the white marker of the image R_A respectively. Since when applying (4) the resulting image I can have intensities greater than 255 or less than zero, an adjustment is made by applying (5) and the final corrected image I_C is obtained.

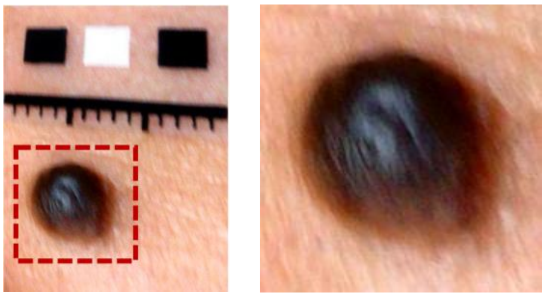
$$I_C = \begin{cases} 255 & , \quad I > 255 \\ I & , \quad 0 \leq I \leq 255 \\ 0 & , \quad I < 0 \end{cases} \quad (5)$$

In Figure 5(a), the image R_A is shown and in Figure 5(b) the enhanced image I_C is shown. This lighting correction process is important to extract the characteristics of the mole correctly. The image is then cropped to just have the mole, without the initial marker. The resulting image I_R is shown in Figure 6.



(a) (b)

Figure 5. Image enhancement process, (a) R image and (b) corrected image I_C



(a) (b)

Figure 6. Image cropping process, (a) input image I_C (b) cropped image I_R

2.4. Segmentation

To the I_R image, the segmentation is made by means of the iterative K-Means grouping method [21], which consists of grouping each pixel according to the smallest distance d, from its centroid c_k, and then recalculating its centroid, this process is repeated until the difference of the centroids is less than 0.001. The flowchart of the K-means segmentation is shown in Figure 7. For this segmentation 2 classes were defined, one class for the mole and one class for the skin. Therefore, the value of k can be, $k = 1, 2$; which results in the image I_K shown in Figure 8.

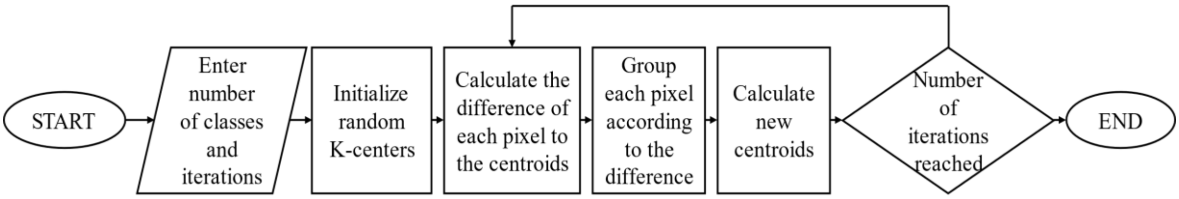

START
Enter number of classes and iterations
Initialize random K-centers
Calculate the difference of each pixel to the centroids
Group each pixel according to the difference
Calculate new centroids
Number of iterations reached
END

Figure 7. K-mean segmentation flowchart

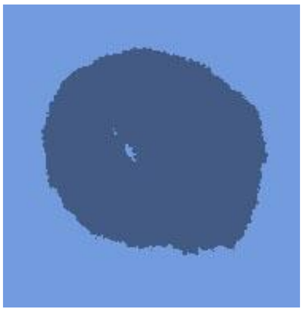


Figure 8. Image I_K showing the 2 classes (two colors)

To have the two regions, you assign the value of 1, to the pixels that belong to the mole group; and you assign 0 to the pixels that belong to the skin group. This is assigned in (6), resulting in the I_S image, shown in Figure 9(a).

$$I_S = \begin{cases} 0 \quad , \quad I_K == skin\ class \\ \\ 1 \quad , \quad I_K == skin\ mole\ class \end{cases} \tag{6}$$

Once the segmentation in the I_S image is obtained with the K-Means algorithm, morphological operations are applied to correct, eliminate imperfections and close possible holes. This process is performed in (7), (8), (9) and (10).

$$IG_1 = (I_S \oplus K) \ominus K \tag{7}$$

$$IG_2 = (IG_1 \oplus K) \ominus K \tag{8}$$

$$IG_3 = (IG_2 \oplus K) \ominus K \tag{9}$$

$$I_B = (IG_3 \oplus K) \ominus K \tag{10}$$

where, K is the structural element shown in (11). Figure 9(b) shows the obtained image I_B.

$$K = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{11}$$

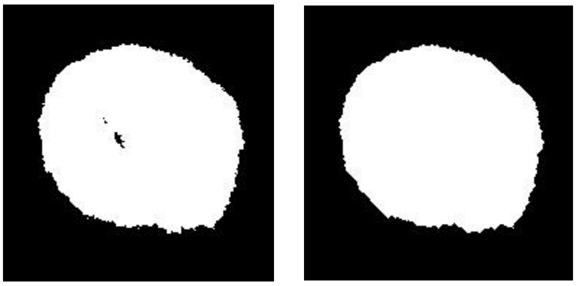


Figure 9. Segmentation result: (a) Image I_S, (b) Image I_B

Obtaining the I_B image, it will serve as a mask to be able to segment the object of interest, which is the mole, this is obtained from the I_R image, this is done by applying (12). The resulting I_F image is shown in Figure 10(c).

$$I_F = I_R \cdot I_B \tag{12}$$

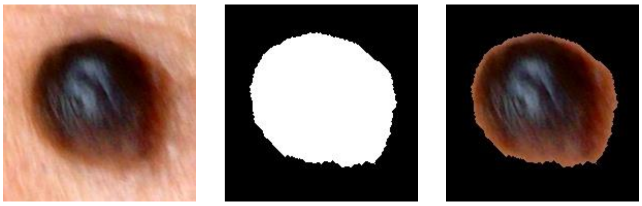


Figure 10. Mole segmentation: (a) The image I_R (b) the image I_B and (c) the segmented mole in the image I_F

2.5. Feature extraction

The proposal made for the analysis of the mole is the extraction of characteristics according to the ABCD method used by the dermatologist for the classification of suspected or non-suspected melanoma. The I_F image of the segmented mole is the image used to perform the feature extraction. The asymmetry will be obtained, then the irregularity of the edges, color and finally the diameter.

2.5.1. Asymmetry

To obtain this characteristic, first the shape of the mole is approximated to an ellipse [22] as shown in Figure 11(a). Once the approximation is obtained, the main axes of the ellipse and the quadrilateral that circumscribes the ellipse are located as shown in Figure 11(b). After this, the image is rotated by taking the circumscribed quadrilateral as shown in Figure 11(c) and finally the image of the mole is cut into four equal parts as shown in Figure 11(d).

For the analysis of the asymmetry a similarity comparison is made of the 4 sub-images obtained from the division of the I_F image, having the upper right side image defined as F_{RU}, the upper left side image defined as F_{LU}, the lower right side image defined as F_{RD} and the lower left side image defined as F_{LD}. These images are shown in Figure 11(d). To analyse the similarity, the logical operation of OR-EXCLUSIVE is performed by applying (13) to each sub-image with its side and bottom and top sides as appropriate.

$$\begin{aligned} C_1 &= F_{LU} \oplus F_{RU} \\ C_2 &= F_{LU} \oplus F_{LD} \\ C_3 &= F_{RD} \oplus F_{RU} \\ C_4 &= F_{RD} \oplus F_{LD} \end{aligned} \tag{13}$$

The images resulting from this operation are C_1, C_2, C_3, y C_4 ; which are the differences that exist between each of the sub-images as shown in Figure 12. From each of these images its area of difference is obtained by adding the number of pixels and dividing by N, which represents the total size of the image, according to (14). These differentiation values will also be used for edge analysis. From the asymmetry analysis, 4 descriptors will be obtained.

$$P_{dif} = \frac{1}{N}\sum_x \sum_y C_i(x,y) \tag{14}$$

where $i = 1,2,3,4$, indicates each of the four sub images.

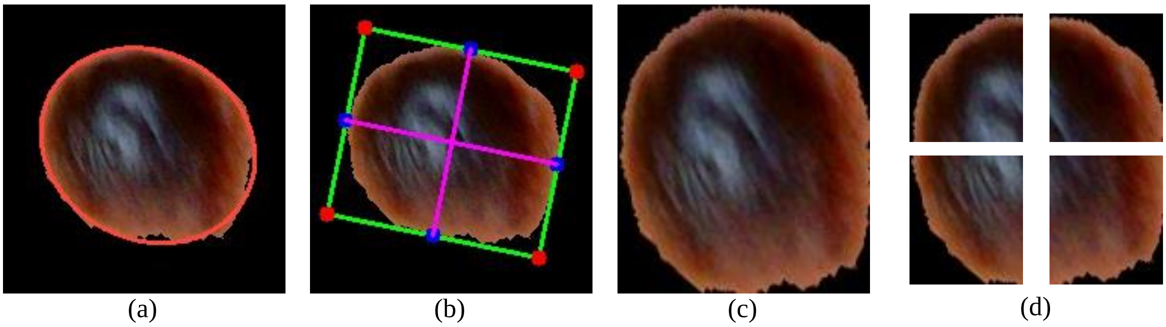


Figure 11. Analysis of the asymmetry: (a) Approximation of the shape to an ellipse, (b) Main axes and quadrilateral circumscribing the ellipse, (c) Image of the rotated mole according to the circumscribing quadrilateral, (d) the four sub-images, upper left side F_{LU}, upper right side F_{RU}, lower left side F_{LD} and lower right side F_{RD}

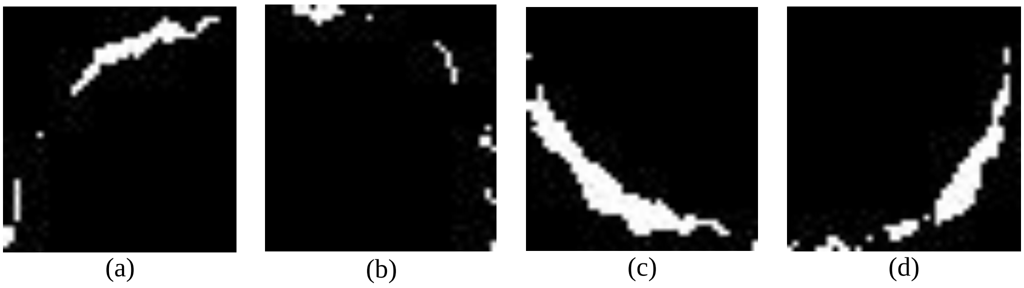


Figure 12. Asymmetry result: (a) C_1, image differentiates between F_{LU} and F_{RU}, (b) C_2, image differentiates between F_{LU} and F_{LD} (c) C_3, image differentiates between F_{RD} and F_{RU}, (d) C_4, image differentiates between F_{RD} and F_{LD}

2.5.2. Border

The second characteristic obtained is the irregularity of the edges that could exist, for this we analyze the edge of the mole. The same procedure as Sancen-Plaza *et al.* in [23] is followed with the area obtained in the process of application (10) shown in Figure 10(b). From this process, 9 descriptors of the degree of irregularity of the border are obtained, according to Santiago-Montero *et al.* [24].

2.5.3. Color

We obtain the mean and variance of each of the color components that make up the images F_{RU}, F_{LU}, F_{RD} y F_{LD} in the RGB, HSV, La*b* and YCrCb color models. Each of the mean and variance values obtained in each sub-image is subtracted with those of the other sub-images in each color model, according to (17) and (18). In (15) is shown the way to calculate the mean E for each of the components where g_{lm} represents the value of each pixel, e l and m are the rows and columns respectively, N the total number of pixels and in (16) is shown the standard deviation σ for each of the components.

$$E = \sum_l \sum_m \frac{1}{N} g_{lm} \tag{15}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_l \sum_m (g_{lm} - E)^2} \tag{16}$$

$$\begin{array}{ll} E1_{dif} = E_{RU} - E_{LU} & E4_{dif} = E_{LU} - E_{RD} \\ E2_{dif} = E_{RU} - E_{RD} & E5_{dif} = E_{LU} - E_{LD} \\ E3_{dif} = E_{RU} - E_{LD} & E6_{dif} = E_{RD} - E_{LD} \end{array} \quad (17)$$

$$\begin{array}{ll} \sigma1_{dif} = \sigma_{RU} - \sigma_{LU} & \sigma4_{dif} = \sigma_{LU} - \sigma_{RD} \\ \sigma2_{dif} = \sigma_{RU} - \sigma_{RD} & \sigma5_{dif} = \sigma_{LU} - \sigma_{LD} \\ \sigma3_{dif} = \sigma_{RU} - \sigma_{LD} & \sigma6_{dif} = \sigma_{RD} - \sigma_{LD} \end{array} \quad (18)$$

where $i = 1,2,3,4$, indicates each of the four sub images.

With the mean and variance in each of the color models we have 12 descriptors that when applied to the 4 models makes a total of 48 descriptors. Additionally, Haralick's textural parameters are used as color characteristics in the RGB color model. According to [25], 14 second order statistical variables are calculated, which describe properties such as contrast, energy, entropy, local uniformity, maximum probability, hue, importance, and correlation to the I_F image. Finally, for the color characteristic there are a total of 62 descriptors.

2.5.4. Diameter

From the image I_B, we obtain the longest straight distance [26], for this we obtain the distance between each of the points of the edge according to (17) where p_s and p_q represent any two points of the edge with coordinates (x_s, y_s) and (x_q, y_q) respectively, then we compare the T distances that the edge of the image could have to find the longest distance according to (18), this value represents the diameter of the mole as shown in Figure 13. This distance is obtained in pixels and divided by 20, because 20 pixels is the equivalent of 1mm, which was obtained using the marker as a reference. In addition, the area is obtained by applying (19) in pixels of the object in the image $I_B(x, y)$ and making the equivalence of 1 mm^2 equivalent to 400 pixels.

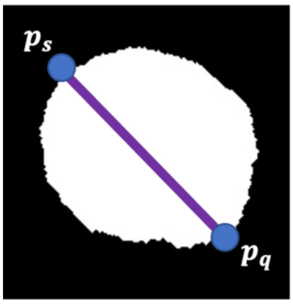


Figure 13. Image I_B showing the diameter of the mole

$$dist(p_s, p_q) = \sqrt{(x_s - x_q)^2 + (y_s - y_q)^2} \quad (17)$$

$$diam = \max\left[dist(p_s, p_q)_1, dist(p_s, p_q)_2, \ldots, dist(p_s, p_q)_T\right] \quad (18)$$

$$area = \sum_x \sum_y I_B \quad (19)$$

where x and y indicate the coordinate of the pixels that make up the image. Likewise, according to [27], the moments invariant to the transformations are used, which are desirable characteristics to recognize the objects more easily. Thus, the 7 Hu moments of the image I_B are obtained invariant to translation, scaling and rotation. Finally, in this section, 9 descriptors are obtained, 2 belonging to the diameter and area and 7 to the moments of Hu.

2.5.5. Metadata

In addition to the characteristics obtained by image processing, data collected by the people in charge of taking the images in the health centers will also be used for the classifier. This information corresponds to the person's age, how long the mole has been there at the time the image was obtained, whether there has been any discomfort, the location of the mole (head, trunk, extremities, palms) and whether there have been any recent changes in the mole. In Table 1, the metadata being used and their respective values corresponding to each of the 10 elements are shown.

Table 1. Values of the metadata

Time: Birth	Time: 1 year	Time: 2-3 years	Time: 4 + years	Pain	Location: Head	Location: extremities	Location: Palms	Location: Chest	Recent changes
1: Yes 0: No	1: Yes 0: No	1: Yes 0: No	1: Yes 0: No	1: Yes 0: No	1: Yes 0: No	1: Yes 0: No	1: Yes 0: No	1: Yes 0: No	1: Yes 0: No

2.5.6. Classifiers

According to [16, 28], the support vector machine (SVM) are supervised learning models, which means that the sample data must be labeled, and can be applied to almost any type of data. It is basically based on the concept of decision planes that separate classes with a hyperplane.

The size of the SVM input vector is 94 and corresponds to the 94 characteristics obtained. In the training phase, the classifier model was built using the cross-validation procedure to find the optimization parameters of the hyperplane to avoid bias with overfitting.

Another classification method used is random forest [17, 29]. In Random Forest, several decision trees are built, instead of just one at the time of training. To classify a new object based on attributes, each decision tree gives a classification and finally the mode of the classes is taken is the output of our classifier.

As in the case of SVM, the size of the input vector for random forest is 94 and corresponds to the 94 characteristics obtained. Finally, Naïve Bayes is a probabilistic classifier [18, 30] that uses Bayes' rule together with a strong assumption that the features are conditionally independent. In this way this classifier can be trained in a supervised way very efficiently, because the number of parameters needed are linear with respect to the number of features of the classes.

3. RESULTS AND ANALYSIS

The dataset consists of 95 photographs of moles taken with a smartphone, by technical personnel in different health centers. In addition, all images have a marker on the side of the mole. In Figures 14 and 15, the results of the mole segmentation process are shown. The result image will be the one that will be analyzed to obtain its characteristics. For the processing and analysis of the images we use the Python programming language, and the OpenCV and scikit-learn libraries.

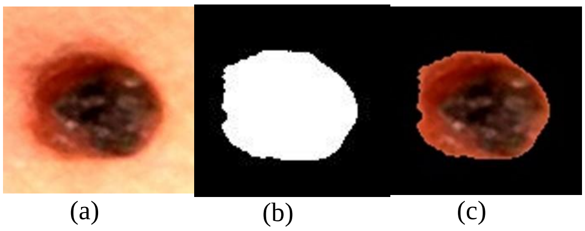


Figure 14. Mole segmentation process (a) The original image of the mole, (b) the binary image, (c) the segmented mole

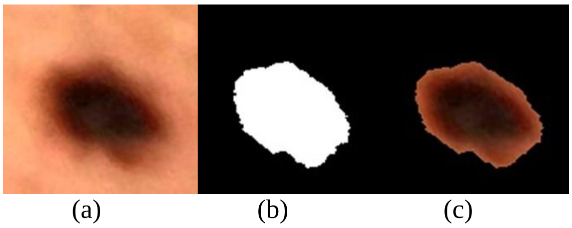


Figure 15. Mole segmentation process (a) the original image of the mole, (b) the binary image, (c) the segmented mole

Of the 95 images processed, 80% of the images are used for training and 20% for the test. The results for the training images, 76 images, can be seen in the confusion matrix in Table 2, where an accuracy of 0.9473, sensitivity of 0.8571 and specificity of 1.0 were obtained. In this case, SVM was used to perform the classification. For the test process, 19 images were used, of which 8 were Not Suspect and 11 Suspect. The accuracy for these data was 0.9473, a sensitivity of 0.909 and a specificity of 1.0. This is shown in Table 2.

Then using another classification method, in this case Random Forest, an accuracy of 0.9868, sensitivity of 0.9668 and specificity of 1.0 was obtained for the training data, showing in Table 3 their respective confounding matrix. And using the same method for the test data, 0.7368 precision was obtained, with a sensitivity of 0.5454 and a specificity of 1.0, shown in Table 3. Using the last classification method, in this case Naive Bayes, an accuracy of 0.7763, sensitivity of 0.7763 and specificity of 0.9166, was obtained for the training data, showing in Table 4 its respective confusion matrix. And using the same method for the test data, 0.6842 precision was obtained, with a sensitivity of 0.6363 and a specificity of 0.75, shown in Table 4.

Table 2. Confusion matrix SVM train and test

		Predicted label (Train)		Predicted label (Test)	
		Not suspicious	Suspicious	Not suspicious	Suspicious
True Label	Not suspicious	48	0	8	0
	Suspicious	4	24	1	10

Table 3. Confusion matrix random forest train and test

		Predicted label (Train)		Predicted label (Test)	
		Not suspicious	Suspicious	Not suspicious	Suspicious
True Label	Not suspicious	48	0	8	0
	Suspicious	1	27	5	6

Table 4. Confusion matrix Naïve Bayes train and test

		Predicted label (Train)		Predicted label (Test)	
		Not suspicious	Suspicious	Not suspicious	Suspicious
True Label	Not suspicious	44	4	6	2
	Suspicious	13	15	4	7

Tables 5 and 6 show the accuracy, sensitivity, and specificity metrics for the different classifiers. Sensitivity indicates the ability of our estimator to identify positive cases and specificity indicates the ability of the estimator to identify negative cases. We observe that specificity values are higher than sensitivity values, so our classifier is better at ruling out the possibility of having melanoma.

Table 5. The performance of classificatory in train

Metric	SVM	Random Forest	Naïve Bayes
Accuracy	0.9473	0.9868	0.7763
Sensitivity	0.8571	0.9642	0.5357
Specificity	1.0000	1.0000	0.9166

Table 6. The performance of classificatory in test

Metric	SVM	Random Forest	Naïve Bayes
Accuracy	0.9473	0.7368	0.6842
Sensitivity	0.9090	0.5454	0.6363
Specificity	1.0000	1.0000	0.7500

4. CONCLUSION

With the training images, for each of the methods, shown in Table 5, we can see that SVM has 0.9473 accuracy compared to random forest that obtains 0.9868, however, for the test images shown in Table 6, the one that obtains better accuracy is SVM with 0.9473 unlike random forest that falls to 0.7368. In addition, the sensitivity for the first classifier is 0.909 which is a big difference to random forest which gets 0.5454. In conclusion, SVM can be taken as the best classifier, even though in training it gave a lower result, in the test data its performance was maintained.

Comparing our results with works that use machine learning techniques as in the cases of Mustafa *et al.,* which obtains 0.8667 and in Murugan *et al.,* obtains 0.8943 using in both cases SVM, we can observe that our results using the same technique are 0.9473. These results support our hypothesis that by using images from different sources to a dermatoscope it is also possible to make the identification of melanoma with great inference power.

In addition, in one of the latest articles using technology assistance to identify skin diseases, such as Liu *et al.*, show that the maximum accuracy obtained by their deep learning based algorithm with dermatoscopic images was 0.90, and in our case it was 0.94 using images taken by smartphones, the results obtained in both studies are similar so it can be concluded that the proposed method of using feature extraction using digital image processing of skin lesions is an effective approach to identify the presence of malignant skin lesions, besides not requiring as many computer resources as it does a deep learning based one. For future work we intend to make use of telemedicine to be able to use our algorithm in populations that do not have access to medical consultations, thus improving health in vulnerable places. We also want to increase our database to be able to have more diversity of samples and to be able to improve our results.

REFERENCES

[1] Hay, Roderick J. *et al.*, "The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions," *Journal of Investigative Dermatology*, vol. 134, no. 6, pp. 1527-1534, 2014.

[2] Rogers, Howard W. *et al.*, "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012," *JAMA dermatology*, vol. 151, no.10, pp. 1081-1086, 2015.

[3] American Cancer Society, "California Cancer Facts & Figures 2017," California Department of Public Health, California Cancer Registry, 2017.

[4] Siegiel, R., K. Miller, and A. Jemal, "Cancer statistics, 2017," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 7-30, 2017.

[5] Brady, Mary S. *et al.*, "Patterns of detection in patients with cutaneous melanoma: implications for secondary prevention," *Cancer*, vol. 89, no. 2, pp. 342-347, 2000.

[6] Kittler, Harold *et al.*, "Diagnostic accuracy of dermoscopy," *The lancet oncology*, vol. 3, no. 3, pp. 159-165, 2002.

[7] "What Are Basal and Squamous Cell Skin Cancers? Types of Skin Cancer," *Cancer.org*, 2020. [Online]. Available: https://www.cancer.org/cancer/basal-and-squamous-cell-skin-cancer/about/what-is-basal-and-squamous-cell.html

[8] “Melanoma Skin Cancer: Understanding Melanoma,” *American Cancer Society*. [Online]. Available: https://www.cancer.org/cancer/melanoma-skin-cancer.html

[9] D. S. Rigel *et al.*, "ABCDE-an evolving concept in the early detection of melanoma," *Archives of dermatology*, vol. 141, no. 8, pp. 1032-1034, 2005.

[10] Y. Liu *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, vol. 26, pp. 900-908. 2020.

[11] Jackson, Robert, "Differential Diagnosis of Some Common Skin Lesions: A Morphological Viewpoint," *Canadian Family Physician*, vol. 22, pp. 78-79, 1976.

[12] M. Emre *et al.*, "A state-of-the-art survey on lesion border detection in dermoscopy images," *Dermoscopy image analysis*, vol. 10, pp. 97-129, 2015.

[13] L. Yu *et al.*, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Transactions on Medical Imagin*g, vol. 36, no. 4, pp. 994-1004, 2016.

[14] E. Nasr-Esfahani *et al.*, "Melanoma detection by analysis of clinical images using convolutional neural network," *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC*), Orlando, FL, USA, 2016, pp. 1373-1376.

[15] S. Mustafa and A. Kimura, "A SVM-based diagnosis of melanoma using only useful image features," *2018 International Workshop on Advanced Image Technology (IWAIT),* Chiang Mai, 2018, pp. 1-4.

[16] A. Murugan, S. A. H. Nair and K. P. S. Kumar, "Detection of skin cancer using SVM random forest and kNN classifiers," *Journal of medical systems*, vol. 43, no. 8, Aug. 2019, Art. no. 269.

[17] M. A. Arasi, E. M. El-Horbaty and E. A. E. El-Dahshan, "Classification of Dermoscopy Images Using Naïve Bayesian and Decision Tree Techniques," *2018 1st Annual International Conference on Information and Sciences (AiCIS),* Fallujah, Iraq, 2018, pp. 7-12.

[18] “Investigadores de la Cayetano crean app que detecta cáncer de piel,” *Dirección de Investigación*. [Online]. Available: https://investigacion.cayetano.edu.pe/articulos-impacto/93-investigadores-de-la-cayetano-crean-app-que-detecta-cancer-de-piel

[19] C. A. Glasbey and K. V. Mardia, “A review of image-warping methods,” *Journal of applied statistics*, vol. 25, no. 2, pp. 155-171, 1998.

[20] S. Collings *et al*, “Non-invasive detection of anaemia using digital photographs of the conjunctiva,” *PloS one*, vol. 11, no. 4, 2016, Art. no. e0153286.

[21] N. Dhanachandra, K. Manglem, Y. JinaChanu, “Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm,” Procedia Computer Science, vol. 54, 2015.

[22] Z. Cheng, Y. Liu, "Efficient technique for ellipse detection using restricted randomized Hough transform," *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.,* Las Vegas, NV, USA, vol. 2, 2004, pp. 714-718.

[23] Sancen-Plaza, A. *et al.*, “Quantitative evaluation of binary digital region asymmetry with application to skin lesion detection,” *BMC medical informatics and decision making*, vol. 18, no. 1, pp. 50-60. 2018.

[24] Santiago-Montero R., Lopez-Morales M. A., Sossa J. H., "Digital shape compactness measure by means of perimeter ratios," *Electronics letters*, vol. 50, no. 3, pp. 171-173, 2014.

[25] R. M. Haralick, K. Shanmugam *et al.*, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621, 1973.

[26] L. Wang, Y. Zhang, and J. Feng, "On the Euclidean distance of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1334-1339, 2005.

[27] Z. Huang and J. Leng, "Analysis of hu's moment invariants on image scaling and rotation," *2010 2nd International Conference on Computer Engineering and Technology*, Chengdu, China, 2010, pp. V7-476-V7-480.

[28] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293-300, 1999.

[29] L. Breiman, "Random Forests," *Statistics Department, University of California, Berkeley*, CA 94720, 2001.

[30] P. Langley, and S. Sage, "Induction of Selective Bayesian Classifiers," *Uncertainty Proceedings 1994*, pp. 399-406, 1994.

BIOGRAPHIES OF AUTHORS

Zaid Sanchez, studied mechatronic engineering at the Universidad Nacional de Ingenieria, Peru (UNI), has developed research projects on topics related to image analysis using digital image processing techniques, Machine Learning and Deep Learning.

Alicia Alva, Master's in biomedical informatics in Global Health, Universidad Peruana Cayetano Heredia in collaboration with the University of Washington Bachelor in Science with a Mention in Mathematics from the National University of Engineering. He has developed in the last years projects of Telemedicine, coordinating, supervising the interventions in the Systems of Telediagnosis of Tuberculosis and Melanoma, with institutions of province in coast and mountain. She is also responsible for the development of diagnostic software based on pattern recognition, for various diseases such as Tuberculosis, melanoma, bacterial vaginosis, intestinal parasitosis, plasmodium falciparum, among others.

Mirko Zimic, degree in Control and Prevention of Diseases from Johns Hopkins University, MSc. Biochemistry from the Peruvian University Cayetano Heredia (UPCH) and obtained his BSc. Physics at the National University of Engineering. Currently, he is the head of the Laboratory of Bioinformatics and Molecular Biology of the UPCH and leads a multidisciplinary working group, with whom he researches in different areas of science and technology.

Christian del Carpio, he received the B.S. degree in Electrical Engineering in 2005 from the San Martin de Porres Private University (USMP), Lima, Peru. In the year 2015, he obtained a master's degree in science, from National University of Engineering (UNI), Peru. He is working as a research professor for the undergraduate and postgraduate programs of the schools of Electrical Engineering and the school of Mechatronics Engineering from the National University of Engineering (UNI), Lima, Peru. His research interests include image and signal processing.