Eco matemático ISSN: 1794-8231 (Impreso), E-ISSN: 2462-8794 (En línea) Volumen 13 (1) Enero-Junio de 2022, páginas 34-42

Preliminary Identification of Skin Lesions using Efficient Computational Learning Techniques

Identificación preliminar de lesiones cutáneas mediante técnicas de aprendizaje computacional eficientes

Carlos Vicente Niño-Rondón[a], Diego Andrés Castellano-Carvajal[b], Sergio Alexander Castro-Casadiego[c], Byron Medina-Delgado[d], Dinael Guevara-Ibarra[e]

[a]Ingeniero Electrónico, Universidad Pontificia Bolivariana Bucaramanga, Colombia, https://orcid.org/0000-0002-3781-4564, carlos.nino.2022@upb.edu.co
[b]Ingeniería Electrónica, Universidad Francisco de Paula Santander, Cúcuta, Colombia, https://orcid.org/0000-0002-4530-1136, diegoandrescc@ufps.edu.co
[c]Magíster en Ingeniería Electrónica, Universidad Francisco de Paula Santander, Cúcuta, Colombia, https://orcid.org/0000-0003-0962-9916, sergio.castroc@ufps.edu.co@ufps.edu.co
[d]Doctorado en Ciencias, Universidad Francisco de Paula Santander, Cúcuta, Colombia, https://orcid.org/0000-0003-0754-8629, byronmedina@ufps.edu.co
[e]Doctor en Ingeniería, Universidad Francisco de Paula Santander, Cúcuta, Colombia, https://orcid.org/0000-0003-3007-8354, dinaelgi@ufps.edu.co

Forma de citar: Niño-Rondón, C. V., Castellano-Carvajal, D. A., Castro-Casadiego, S.A., Medina-Delgado. B., Guevara-Ibarra D. (2022). Preliminary Identification of Skin Lesions using Efficient Computational Learning Techniques. *Eco Matemático,* 13 (1), 34-42

Recibido: 26 de mayo de 2021
Aceptado: 8 de septiembre de 2021

Palabras clave

Skin Lesions,
Feature Extraction,
Computational Learning,
Open-Source Tools.

Abstract: Machine learning (ML) is one of the fields of artificial intelligence that offers algorithms to predict from samples the effective detection of skin lesions caused by skin cancer. This paper presents the preliminary identification of skin lesions using optimized algorithms for texture feature extraction by GLCM and feature-based learning (LightGBM, SVM and HAAR Cascade) as an initial stage for a diagnostic tool. The HAM10000 skin lesion image set, Python programming language and open source-based libraries are used to process the images, extract the features and train the learning models, determine the performance and hit rate of the models. Based on the results obtained, the LightGBM classifier required the shortest learning time, reduced CPU usage and 90 % accuracy rate.

Keywords

Lesiones Cutáneas,
Extracción de Características,
Aprendizaje Computacional,
Herramientas de Código Abierto.

Resumen: El aprendizaje automático (ML) es uno de los campos de la inteligencia artificial que ofrece algoritmos para predecir a partir de muestras la detección efectiva de lesiones cutáneas causadas por el cáncer de piel. En este trabajo se presenta la identificación preliminar de lesiones cutáneas utilizando algoritmos optimizados para la extracción de características de textura mediante GLCM y aprendizaje basado en características (LightGBM, SVM y HAAR Cascade) como etapa inicial para una herramienta de diagnóstico. El conjunto de imágenes de lesiones cutáneas HAM10000, el lenguaje de programación Python y las bibliotecas basadas en código abierto se utilizan para procesar las imágenes, extraer las características y entrenar los modelos de aprendizaje, determinar el rendimiento y la tasa de aciertos de los modelos. Según los resultados obtenidos, el clasificador LightGBM fue el que requirió el menor tiempo de aprendizaje, un uso reducido de la CPU y una tasa de acierto del 90 %.

*Autor para correspondencia: carlos.nino.2022@upb.edu.co

Doi: https://doi.org/10.22463/17948231.3286

Introduction

Skin cancer is considered to be the most frequent disease in the world with incidence rates following a progressive increase during the last decade, affecting mainly fair-skinned populations. Non-melanoma skin carcinoma and squamous cell carcinoma represent the majority of skin pathologies. Globally, the incidence of melanoma follows an annual trend of 4 % to 5 %, and by 2018, there were almost 126,000 deaths from skin cancer (Parker, 2021). Machine learning (ML)-oriented artificial intelligence (AI) algorithms for skin cancer analysis have augmented a number of strategies for detection using simple and deep learning-based algorithms implemented to evaluate skin lesions by biomedical imaging of datasets with promising classification results (Magalhaes et al., 2021)

On the other hand, gray-level co-occurrence matrix (GLCM) is presented as a method applied to feature extraction on lesion pattern as a measure for texture representation, which has demonstrated the ability to distinguish different types of skin cancer with an accuracy of better than 76 % (Chandra et al., 2019). LightGBM refers to an ensemble-based learning technique that trains under a decision tree with gradient boosting and provides a robust model composed of weak classification models that largely eliminates model fitting for good results (Ahmed et al., 2020).

The support vector machine (SVM) is one of the learning algorithms used in data classification, based on decision planes that separate objects in order to differentiate between different classes that are used to classify the image into categories. SVMs have demonstrated with experimental results accuracy higher than 77 % (Almansour & Arfan, 2016). Similarly, trained cascade classifiers are a tool used for validation and recognition of melanoma in medical images that help clinicians and reduce the mortality rate by recognizing items. The results show a superior specificity of 90.05 % and sensitivity of 83.06 % due to cross-validation and application of the scanning method and statistical evaluation on the image (Afifi et al., 2017).

This paper presents a comparative study aimed at the initial identification of skin lesions in dermoscopic images by applying LightGBM classifiers combined with GLCM for feature extraction, support vector machine and HAAR cascade classifiers based on machine learning (ML). The robustness of the classifiers is validated by quantifying the training time of the system and the requirements of the Central Processing Unit (CPU) used. The coding developed has been based on the Python programming language with the utility of Pandas, OpenCV, Numpy and Sklearn libraries for computer vision, data analysis and classifier training. The selected test images are part of the HAM10000 dataset.

Related Works

Almansour & Arfan Jaffar present the combination of a GLCM algorithm with LBP to extract energy, contrast, entropy, homogeneity and LBP matrix features to classify skin anomalies with an accuracy of 90.32 % and sensitivity of 85.84 % (Almansour & Arfan, 2016). Afifi, Gholamhosseini & Sinha present the development of a low-cost biomedical device on FPGA platform that performs real-time diagnostics based on SVM and cascade classifiers for timely detection of melanoma. Hardware implementation results demonstrated an accuracy of 97.90 %, software acceleration factor of 26, resource utilization of 34 % and 2 watts in power consumption (Afifi et al., 2017). (Li et al., 2018) proposed the LightGBM classifier to combine and select salient features for deep learning model parameter tuning in the representation of clinical criteria focused on skin lesion diagnosis tasks.

Methodology

An applied and comparative methodology based on three stages is proposed. In the first stage, acquisition, grayscale transformation and texture

feature extraction of the training and evaluation images are performed. In the second stage, training with LightGBM learning techniques, support vector machine and cascaded classifiers is performed with the training images to determine the accuracy. Furthermore, in the third stage, the response time, throughput and performance of each learning algorithm is validated with the different test images.

Image Preprocessing and Feature Extraction

The set of images used for training and evaluation are part of the HAM10000. The set of images used for training and evaluation are part of the HAM10000. The extraction of texture features requires preprocessing in grayscale transformation using intensity values between shades of black and white (Padmavathi & Thangadurai, 2016). So, equation 1 presents the expression in weighted sums of the R, G and B components respectively for grayscale conversion.

$$I=0.299R+0.587G+0.114B \quad (1)$$

The evaluation of texture-based features of the image is achievable with the second-order statistical technique GLCM (gray level co-occurrence matrix) which are calculated from metrics with angular ratios and distances between pixels (Kumar et al., 2020). The matrix solution is based on reducing the number of gray values and the number of pixel combinations. In addition, it is necessary to define the direction of the pixel pair and the distance between them, since the matrix stores the number of times a pair is found in the image and, at the same time, analyzes the relationship between non-consecutive pixels, so the distance between pixels is defined beforehand (Gajbar & Deshpande, 2015). Therefore, Figure 1 shows in (A) the matrix with the number of occurrences between adjacent pixels, and in (B) the GLCM displacement directions.

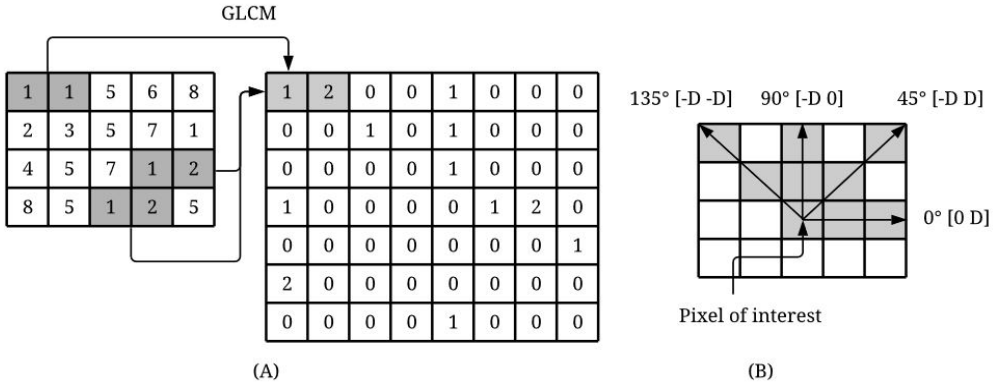

Figure 1. Construction process of a GLCM matrix.

Similarly, from the gray level co-occurrence matrix, second order textural variables are calculated by pixel analysis starting from a given distance and angular orientations (Ríos et al., 2009). The texture-based variables are presented in Table I.

Table I. Variables Related to GLCM Textures

GLCM Feature	Explanation	Model
Energy	Measures uniformity in relation to image texture.	$\sqrt{\sum_{i,j=0}^{N-1} C_{i,j}{}^2}$
Correlation	It measures the dependence of linear amplitudes on neighboring amplitudes. The variables σ_i and σ_j are the standard deviations, while, μ_i and μ_j are the GLCM averages.	$\sum_{i,j=0}^{N-1} \frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2 * \sigma_j^2}}$
Dissimilarity	The weights of the probabilities increase linearly and move away from the diagonal where the neighboring values are identical.	$\sum_{i,j=0}^{N-1} C_{i,j} * \lvert i-j \rvert$
Homogeneity	It is responsible for weighting the values by the inverse of the contrast weight. The homogeneity increases as the contrast between pixels decreases.	$\sum_{i,j=0}^{N-1} \frac{C_{i,j}}{1+\lvert i+j \rvert}$
Contrast	Presents the variations over the shades of gray of the image. If the variation of shades of gray is greater, the contrast will be high. The contrast is null if the gray levels are constant.	$\sum_{i,j=0}^{N} (i-j)^2 * C_{i,j}$

Application of computational intelligence techniques

LightGBM is a machine learning classification technique based on an improved gradient boosting framework for decision trees in order to increase model efficiency and reduce memory usage rate. The algorithm uses the gradient-based one-sided sampling (GOSS) technique and exclusive feature clustering (EFB) that complement the histogram used by GBDT (Gradient Boosting Decision Tree) frameworks that group feature or attribute values into discrete garbage cans which accelerates training (Zhang et al., 2019). Most techniques focused on decision tree learning grow the tree horizontally (depth). In contrast, LightGBM grows the tree in a leaf shape, allowing to choose the leaf with maximum delta loss to grow, and by keeping this algorithm fixed by leaves, lower loss is achieved compared to level set algorithm (Chen et al., 2019). Figure 2 illustrates a diagram of the growth of the tree leaves.

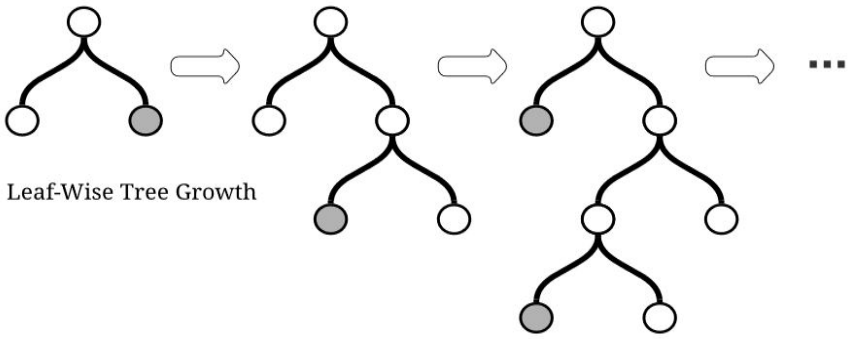

Figure. 2. Growth of LightGBM learning sheets.

The configuration for the creation of the LightGBM classifier model was set with an increased learning rate for training of 0.05, gradient boosting supported with Gradient Boosting Decision Tree, multiclass

learning task, LGBM Regressor evaluation metric, 100 maximum tree leaves for basic learning, maximum learning depth set by default and three iterations for training.

The support vector machine (SVM) has the purpose of finding a hyperplane that is defined within a dimension space based on the number of features for classification of representative data points. Class separation generates several selectable hyperplanes for maximum distance representation between data points to obtain classification with higher accuracy (Lingaraj et al., 2021). The decision level or hyperplane equation separating the input space classes is defined as shown in equation 2, where w^T defines the hyperplane of optimal separation and b is the slope.

$$w^T x_i + b = 0 \qquad (2)$$

The dimension of the hyperplane is related to the number of features and the support vectors. The latter are the points that are closer to the hyperplane and influence the orientation and location of the hyperplane that maximizes the margin of the classifier (Cervantes, 2020). For training the SVM model according to the trend of training data, the parameter gamma was set in the interval between 0.0001 and 1 defining the influence of training examples, C between 0.1 and 100 to compensate the effective classification of training examples, and the kernel selectable between radial basis function and polynomial function.

On the other hand, the cascade classifier performs the filtering successively to the image with various scales and positions considering the difference taken by the pixels as the analysis feature and focuses on the change of value they take on to the image as developed with gradients (Srinivasan & Srinivasna, 2020). In addition, the value of the features with edge and line relation is represented in equation 3, where k is the number of rectangles, μ^i is the average intensity of the pixels present of the rectangle and w^i is the weight of each rectangle.

$$f = \sum_{\dot{\varepsilon}=1}^{k} w^{(i)} * \mu^{(i)}$$

Therefore, Figure 3 shows the binarized HAAR rectangles that are subtracted with the value of the black and white rectangle.

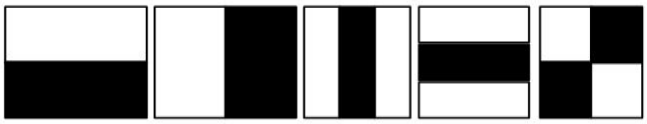

Figure 3. HAAR-type features.

The Cascade Trainer GUI application was used to establish the configuration to obtain the classifier model by initially loading the positive (lesions) and negative (skin) images and defining the percentage of images to be used for training, in this case 100 %. In addition, the number of stages and threads is configured, being 20 and 5 and the dimensions of the samples of 75x75 establishing the type of HAAR features. For the impulse of the model, GAB is defined in the Gentle Adaboost algorithm, and the minimum hit rate is set to 0.995.

Validation

For validation of the proposed feature extraction and learning methods for initial skin lesion identification, an inspection of the hardware tool task manager is performed to determine the learning performance of each proposed method. In addition, the learning and classification techniques are compared using hit rates to visualize the effectiveness of each learning algorithm.

Results and Discussion

According to the proposed methodology, a collection of 456 images was taken for training the

learning models, corresponding to 171 images of skin lesions and 285 images of skin sections without any affectation. For validation, 10 additional images different from the training images were used, corresponding to 5 images of skin lesions and 5 images of skin sections without abnormality. In addition, the images used for training and validation were transformed to grayscale, where, Figure 4 shows in (A) the test images used from the HAM10000 of both lesions and skin and in (B) the grayscale representation of the images.

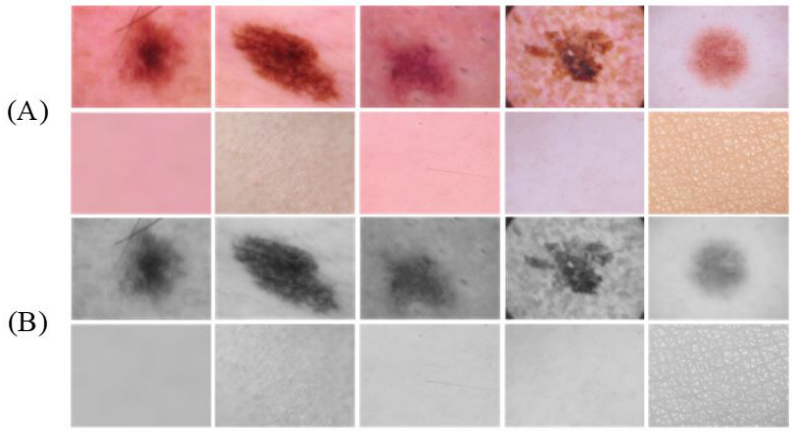

Figure 4. Selected sample images from HAM10000. **Source** [17].

Similarly, representative GLCM texture feature variables were calculated for each of the 456 labeled images selected for LightGBM classifier training and the 10 images selected for evaluation of the three proposed learning methods. Table II shows the features for the 10 sample images for evaluating the LightGBM classifier.

Table II. GLCM Texture Variables Obtained From The Sample Images.

Image type	Energy	Correlation	Dissimilarity	Homogeneity	Contrast
Lesion	0.0576	0.9879	2.9108	0.3384	18.2095
Lesion	0.0513	0.9935	3.2395	0.3651	25.2247
Lesion	0.0563	0.9873	2.3470	0.3576	10.4396
Lesion	0.0323	0.9556	5.8400	0.1841	70.0843
Lesion	0.0957	0.9902	1.6775	0.4924	6.5092
Skin	0.4244	0.9714	0.2335	0.8884	0.2854
Skin	0.0675	0.8754	2.5142	0.3455	11.5391
Skin	0.1814	0.6965	1.1888	0.5367	2.8541
Skin	0.1349	0.9581	1.1749	0.5472	2.6622
Skin	0.0470	0.6011	5.3523	0.1979	52.5980

On the other hand, the time used for training the LightGBM learning method was 1 minute, while the support vector machines required 40 minutes and 10 seconds to train. Likewise, the time used by the cascade classifier system for training was 40 minutes and 20 seconds. Regarding the performance metrics in terms of accuracy, the LightGBM classifier presented an accuracy of 90 % in lesion and skin detection in the images,

the SVMs presented an accuracy of 100 % and, in the cascade classifiers, the accuracy obtained was 100 %. Figure 5 shows the predictions of the proposed learning techniques. For the LightGBM model, out of 5 images of skin lesions, the effective estimation of all 5 images was achieved. However, out of the 5 images of skin sections, only 4 were effectively classified. Likewise, when evaluating the SVMs with the same test images, an effective classification between classes was presented, and when applying the cascade classifiers, an accurate prediction was achieved for both the 5 images with lesions and the 5 images of skin sections.

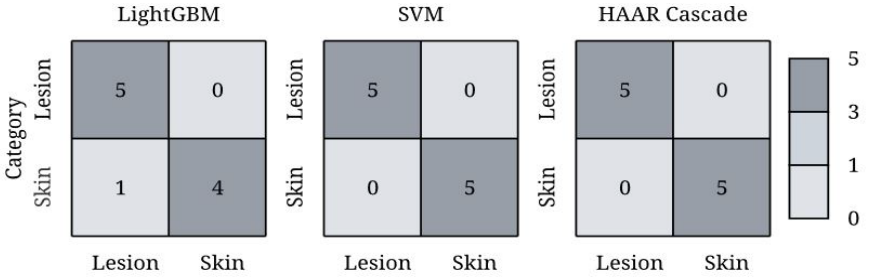

Figure 5. Predictions made by the proposed classifiers.

Figure 6 shows the validation by requirement usage of the AMD Ryzen 5 4500U CPU to execute instruction sequences and process data for training each of the proposed computational learning techniques. As shown, the average CPU usage required to train the LightGBM classifier was 16.50 %. While, for SVM, 41 % of the machine resources were required, and for the cascaded classifiers, the average CPU utility value was 54.25 % respectively.

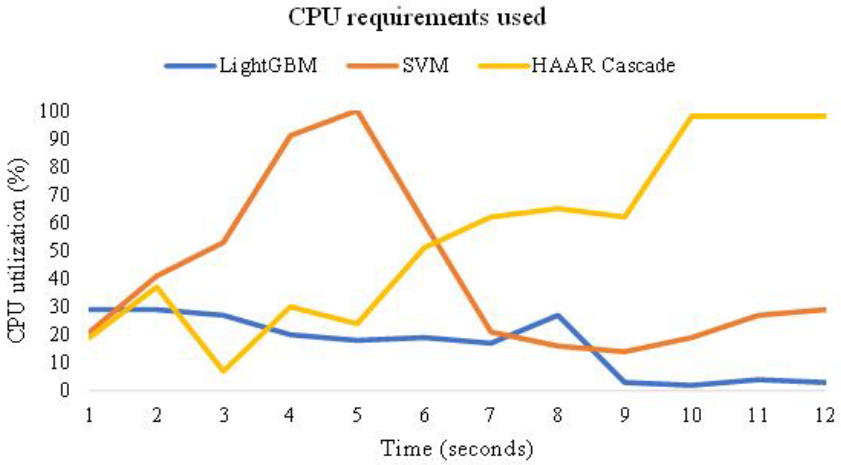

Figure 6. CPU usage of each classifier in training.

Conclusions and future work

The applied methodology allowed identifying the performance and robustness of each proposed classification technique in terms of feature learning and training time, machine resource usage and prediction accuracy rates. Although SVMs and cascade classifiers, according to the results, presented the highest accuracy, they require more time and CPU resources for feature learning of each image in the sample set. In contrast, the LightGBM technique with GLCM for feature extraction presented 90 % accuracy, required one minute for training and the CPU usage was approximately 3 times lower compared to SVMs and cascaded classifiers, making it a replicable technique on high-

end, low-cost hardware tools such as the Raspberry Pi board. Also, although the research focused on preliminary detection of skin lesions, the developed method allows its expansion towards supporting assisted diagnosis of skin diseases, considering the preservation of texture-based features present in the image.

References

Afifi, S., Gholamhosseini, H. & Sinha, R. (2017). SVM classifier on chip for melanoma detection. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, 270–274, doi: 10.1109/EMBC.2017.8036814.

Ahmed, S.A., Yanikoglu, B., Goksu, O., & Aptoula, E. (2020). Skin Lesion Classification with Deep CNN Ensembles. *2020 28th Signal Process. Commun. Appl. Conf. SIU 2020 - Proc*., Oct. 2020, doi: 10.1109/SIU49456.2020.9302125.

Almansour, E. & Arfan Jaffar, M. (2016). Classification of Dermoscopic Skin Cancer Images Using Color and Hybrid Texture Features. *IJCSNS Int. J. Comput. Sci. Netw. Secur*., 16 (4), sp.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. & López, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408, 189–215, doi: 10.1016/J.NEUCOM.2019.10.118.

Chandra, T.G., Nasution, A.M. & Setiadi, I.C. (2019). Melanoma and nevus classification based on asymmetry, border, color, and GLCM texture parameters using deep learning algorithm. *AIP Conference Proceedings*, 2193 (1), 1–6, doi: 10.1063/1.5139389.

Chen, C., Zhang, Q., Ma, Q. & Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemom. Intell. Lab. Syst*., 191, 54–64, doi: 10.1016/J.CHEMOLAB.2019.06.003.

Gajbar A. M. & Deshpande, A. (2015). GLCM and Multiclass Support Vector Machine Based Automatic Detection and Analysis of Types of Cancer and Skin Allergy, *Int. J. Adv. Res. Electron. Commun. Eng*., 4 (5), 1477–1488,

Kumar, M., Alshehri, M., AlGhamdi, R., Sharma, P. & Deep, V. (2020). A DE-ANN Inspired Skin Cancer Detection Approach Using Fuzzy C-Means Clustering. *Mob. Networks Appl*, 25 (4), 1319–1329, doi: 10.1007/S11036-020-01550-2.

Li, X., Wu, J., Jiang, H., Chen, E. Z., Dong, X. & Rong, R. (2018). Skin Lesion Classification Via Combining Deep Learning Features and Clinical Criteria Representations. *bioRxiv*, 1–7, doi: 10.1101/382010.

Lingaraj, M., Senthilkumar, A. & Ramkumar, J. (2021). Prediction of Melanoma Skin Cancer Using Veritable Support Vector Machine. Ann. Rom. Soc. *Cell Biol*., 25, 2623 – 2636.

Magalhaes, C., Tavares, M. R., Mendes, J & Vardasca R. (2021). Comparison of machine learning strategies for infrared thermography of skin cancer. *Biomed. Signal Process. Control*, 69, 1–10, doi: 10.1016/j.bspc.2021.102872.

Padmavathi K. & Thangadurai, K. (2016). Implementation of RGB and Grayscale Images in Plant Leaves Disease Detection – Comparative Study. *Indian J. Sci. Technol*., 9 (6), 1–6, doi: 10.17485/IJST/2016/V9I6/77739.

Parker, E.R. (2021). The influence of climate change on skin cancer incidence – A review of the evidence. *Int. J. Women's Dermatology*, 7

(1), 17–27, doi: 10.1016/J.IJWD.2020.07.003.

Ríos, J., Payá Martínez, J. & Del Baño Aldedo, M. E. (2009). El análisis textural mediante las matrices de co-ocurrencia (GLCM) sobre la imagen ecográfica del tendón rotuliano es de utilidad para la detección de cambios histológicos tras un entrenamiento con plataforma de vibración, *Cult. Cienc. y Deport*., 4 (11), 91–102.

Srinivasan, P. & Srinivasna, V. (2020). A Comprehensive Diagnostic Tool for Skin Cancer Using a Multifaceted Computer Vision Approach, *7th Int. Conf. Soft Comput. Mach. Intell. ISCMI 2020*, 213–217, doi: 10.1109/ISCMI51676.2020.9311557.

Tschandl, P., Rosendahl, C. & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, *Sci. Data 2018 51*, 5(1), 1–9, doi: 10.1038/sdata.2018.161.

Zhang, J., Mucs, D., Norinder, U. & Svensson, F. (2019). LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets. *J. Chem. Inf. Model*., 1–9, doi: 10.1021/ACS.JCIM.9B00633/SUPPL_FILE/CI9B00633_SI_001.PDF.