UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Application of Deep Learning CNN Models for the Detection and Classification of Melanoma Cancer.

José Martín Cadena Zapata

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 07 de febrero de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Hoja de Calificación

de Trabajo de Fin de Carrera

Application of Deep Learning CNN Models for the Detection and Classification of Melanoma Cancer

José Martín Cadena Zapata

Nombre del profesor, Título académico **Noel Pérez Pérez, PhD.**

Quito, 07 de febrero de 2023

© DERECHOS DE AUTOR

Nombres y apellidos: José Martín Cadena Zapata

Código: 206243

Cédula de identidad: 1719959296

Lugar y fecha: Quito, 07 de febrero de 2023

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

El melanoma canceroso es una lesión cutánea relativamente rara que, de ser detectado, puede causar la muerte de una persona debido a su alta tasa de mortalidad. La producción excesiva de melanocitos causa melanoma canceroso en la piel debido a la alta exposición a la radiación solar y al mal cuidado de la piel frente a estas condiciones. Por esta razón, decidimos utilizar modelos de aprendizaje profundo para ayudar detectar melanoma sin necesidad de extraer muestras de piel para biopsias. Para esto, propusimos un nuevo modelo de aprendizaje profundo llamado CNN-2, basado en una arquitectura de red neuronal convolucional para clasificar con éxito las lesiones cutáneas en un conjunto de datos de 2860 lesiones cutáneas tomado del Archivo ISIC. El modelo CNN-2 fue entrenado junto con su modelo base CNN-1 en las mismas condiciones, utilizando un esquema de validación cruzada de 10 veces estratificado, obteniendo un resultado de AUC 0.902 ± 0.03 en el entrenamiento del mejor modelo y un AUC de 0.960 en la prueba para su generalización. Este CNN-2 modelo permitió distinguir entre lesiones cutáneas benignas y melanoma, con características diferentes.

Palabras clave: Red Neuronal Convolucional, Validación Cruzada Estratificada de k-fold, Aprendizaje Profundo, Melanoma, Lesión Cutánea, Clasificación

ABSTRACT

Cancerous melanoma is a relatively rare skin lesion that, if detected, can cause the death of a person due to its high mortality rate. The excessive production of melanocytes causes cancerous melanoma in the skin due to high exposure to solar radiation and poor skin care against these conditions. For this reason, we decided to use deep learning models to help detect melanoma without the need to extract skin samples for biopsies. For this, we proposed a new deep learning model called CNN-2, based on a convolutional neural network architecture to successfully classify skin lesions over a dataset of 2860 skin lesions taken from the ISIC Archive. The CNN-2 model was trained together with its base model CNN-1 under the same conditions, using a stratified 10-fold cross-validation scheme, obtaining a result of AUC 0.902 ± 0.03 in the training of the best model and an AUC of 0.960 in the test for its generalization. This CNN-2 model made it possible to distinguish between benign skin lesions and melanoma, with different characteristics.

Key words: CNN, stratified k-fold cross-validation, Deep Learning, Melanoma, Skin lesion, Classification

TABLA DE CONTENIDO

Introduction **10**

Materials and Methods **14**

Database 14

Deep Learning Models 14

Proposed Method 15

Experimental Setup 17

Data processing 17

Training, validation and test sets 17

Model Configuration 17

Assessment metrics 18

Selection criterion 18

Results and Discussion **19**

Performance evaluation in the training set 19

Performance evaluation in the test set 21

Conclusions and Future Work **23**

Acknowledgment **24**

References **25**

Anexo A: Source Code - GITHUB **28**

ÍNDICE DE TABLAS

Table I. Performance results of deep learning models ……..19

ÍNDICE DE FIGURAS

Figure 1. Examples of image samples available within the database: benign skin lesions (first row) and melanoma diagnosed lesions (second row) ……14

Figure 2. CNN-1 Proposed Architecture ……16

Figure 3. Mean of Loss vs. Epochs ……...20

Figure 4. To the left, ROC-AUC Curve. To the right, Precision vs. Recall Curve ……21

Figure 5. Some samples of the test classification results: columns 1 & 2 were benign and classified as benign, columns 3 & 4 were melanoma and classified as melanoma, and column 5 was benign and classified as melanoma. Green represents and accurate classification. Red represents a wrong classification ……..22

INTRODUCTION

Malignant (cancerous) melanoma is a type of tumor caused by the massive increase and malignant transformation of melanocytes in the skin. Melanocytes are the cells responsible for giving color to the skin; for this reason, malignant melanoma is a disease that occurs, with great tendency, in people whose skin is fair [1]. Although it is a rare dermatologic cancer, its mortality is relatively high. According to [2], cancerous melanoma is responsible for 80% of deaths from skin cancer, which is why it is a subject that is constantly under investigation for its care and prevention.

In the case of Ecuador, diagnoses of cancerous melanoma have increased considerably in recent decades. However, it is not possible to keep real statistics on the current situation in the country, due to the limitations of the National Tumor Registry, whose focus is on the main cities, such as Quito, Guayaquil, Cuenca, and Manta [3]. Due to the geographical location of the country, Ecuadorian citizens receive a large amount of perpendicular solar radiation per year, causing Ecuador to be one of the 20 countries that report the most cancerous melanoma cases annually. Moreover, as the occurrence of cancerous melanoma is linked to sun exposure of people, melanoma rates in Ecuador occur at all ages, for example, from the age of 20 years for the inhabitants of the highland region [4], and in difficult-to-treat areas such as facial regions [5].

Therefore, given the importance of melanoma worldwide, different methodologies for its rapid and accurate diagnosis have been analyzed during the last decade. Some of these methodologies take into account the implementation of Machine Learning (ML) models for the classification of medical images. Shallow and deep learning classifiers have been proposed, depending on how the data are interpreted. Arasi et al. [6], for example, compared two ML models for the binary classification of skin lesions as melanoma or non-melanoma. While Arasi et al. [6] extracted significant features from 206 dermatoscopic images based on Discrete Wavelet Transform (DWT) and texture analysis with Gray Level Cooccurrence Matrix

(GLCM). A Naive Bayes classifier performed the best with 98.8% accuracy, 97.9% sensitivity, and 100% specificity. Comparative results showed that this method has higher accuracy than other state-ofthe-art methods. Almaraz-Damian et al. [7] extracted 15 features from 206 dermatoscopic images using Asymmetrical, Border, Color, Diameter (ABCD) diagnostic criteria, and texture analysis. Almaraz-Damian et al. [7] formulated different feature vectors and fed them to a linear SVM for classification. The best results were obtained with a feature vector of 15 features, with an accuracy of 79.8%.

Regarding image processing, the state of art shows that the first choice is to use of deep learning models, especially, transfer learning. For example, Sagar and Jacob [8] explored the feasibility of applying transfer learning for skin lesion classification. The analysis was performed on 3,000 images from the International Skin Imaging Collaboration. The best result was obtained with ResNet50, with an accuracy of 0.935, precision of 0.94, recall of 0.77, F1 score of 0.85, and ROCAUC of 0.861. Although good results were obtained, a dataset containing more samples per class is expected to improve the results, especially for melanoma. Sallam et al. [9] compared the application of pre-trained CNN models for cancerous melanoma classification using the International Skin Imaging Collaboration 2019 Challenge dataset, which consists of 10,275 images, of which 4,275 represent melanomas. The best result was obtained by GoogleNet, with an accuracy of 0.902 and a validation loss of 0.24. These results are plausible when compared to other models used in this type of classification problem. Jojoa et al. [10] proposed a skin lesion classification system based on Convolutional Neural Networks (CNN). A mask- and region-based CNN is used for segmentation, and ResNet152 for the classification of 2000 dermatoscopic images from the International Skin Imaging Collaboration (ISIC) 2017 Challenge. An accuracy of 0.904, a balanced accuracy of 0.872, a sensitivity of 0.820, and a specificity of 0.925 were obtained. The model can discriminate between benign and malignant lesions. Song et al. [11] proposed a method of integrating five pre-trained Deep CNN (DCNN)

models for feature extraction from 2000 dermoscopic images. A new locally connected neural network was created for classification. The results present an accuracy of 0.909, precision of 0.859, recall of 0.808, f1 score of 0.828, and AUC of 0.911. The integration of trained and optimized DCNNs proves to be applicable for image classification. Jiahao et al. [12] proposed a dermoscopic image classification system for the diagnosis of malignant melanoma. Unlike other researchers, Jiahao et al. [12] used Efficient-B5, a pre-trained DCNN, for feature extraction and classification. A ROC-AUC of 0.919 was obtained. This model is shown to achieve better results than other popular melanoma classifiers.

On the other hand, and due to the scarcity of labeled data in medicine, Pham et al. [13] proposed a method using data augmentation techniques for training the DCNN. The model was trained on 2000 dermatoscopic images and achieved the best results with a neural network. This model achieved an accuracy of 0.89, a specificity of 0.97, a sensitivity of 0.56, and an AUC of 0.892. This shows that medical image classification can benefit from data augmentation. Another example of data augmentation is used by Nasr-Esfahani et al. [14], who focused on pre-processing a set of 170 clinical images and then classifying them with a pre-trained CNN. Preprocessing takes care of reducing photo artifacts such as noise and highlights. The model reaches an accuracy of 0.81, an NPV of 0.86, a PPV of 0.75, a specificity of 0.80, and a sensitivity of 0.81 on clinical images. This method leaves the feature extraction process to the CNN. The most recent approach was proposed by Adegun and Viriri [15] and consisted in using an encoder-decoder network method for feature extraction and a softmax classifier for classification over a dataset that contains 2000 images from the International Skin Imaging Collaboration (ISIC) Archive. An accuracy of 0.95, a dice score of 0.92, a sensitivity of 0.97, and a specificity of 0.96 were achieved. This method aims to eliminate the problems of inhomogeneous features and fuzzy boundaries of the images of skin lesions. Considering the background of this section, it has been shown that the classification of cancerous melanoma is

a topic of relevance in the field of bioinformatics. For this reason, we propose the use of a CNN for the classification of melanoma, in order to achieve better accuracy for the reliable diagnosis of this type of cancer.

MATERIALS AND METHODS

Database

This work considered the use of skin lesion images from the publicly available International Skin Imaging Collaboration (ISIC) Archive [16]. ISIC is a partnership between academia and industry designed to facilitate the application of digital skin imaging to help reduce melanoma mortality. The ISIC Archive serves as a public imaging resource for the development and testing of diagnostic artificial intelligence algorithms. The ISIC Archive contains more than 150,000 images in total, of which approximately 70,000 have been made public.

The experimental dataset consists of 2860 images taken from the ISIC Archive, where half corresponds to melanoma diagnosis, and the other half was diagnosed as benign skin lesions. Figure 1 shows an example of the skin lesions to be worked on within this research. Although the diagnosis of melanoma is small compared to other skin lesions, the generation of a balanced dataset will give the model the ability to be unbiased with respect to a benign lesion.

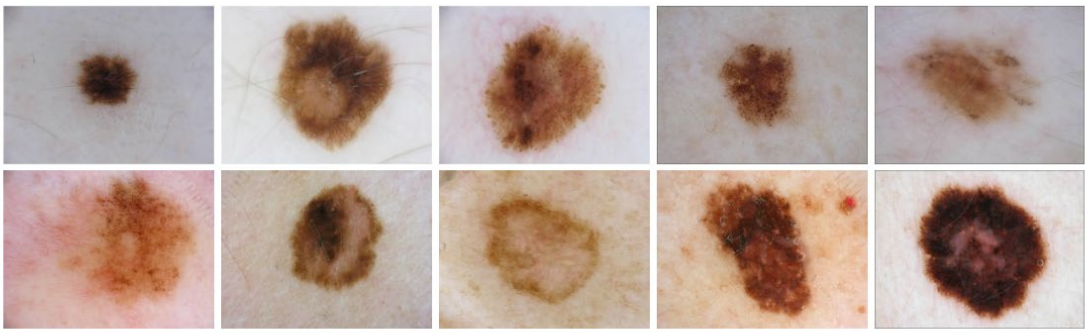

Fig. 1: Examples of image samples available within the database: benign skin lesions (first row) and melanoma diagnosed lesions (second row).

Deep Learning Models

Deep Learning is a specific subfield of machine learning that draws its architecture from the structure and functioning of a brain. In contrast to conventional machine learning, deep learning algorithms are described as ”layered representations learning” [17]. Deep learning algorithms

consist mostly of neural networks, and their effectiveness has been increasing in recent years, as have their applications. For example, deep learning has achieved image classification, speech recognition, handwriting transcription, and even more accurate translations. Its popularity stems from the ability to overcome the performance of machine learning models when they have reached a limit that cannot be surpassed, even by increasing the amount of data to train these models. [18]

For this research, we will use a CNN. A CNN is a Deep Learning model inspired by how the cortex of the human brain works to recognize objects [17]. For this reason, it is one of the preferred choices for image processing. CNNs are usually composed of convolutional layers, pooling (or subsampling) layers, and a fully connected layer at the end. The convolutional layer is composed of a series of filters that are applied to all areas of an input. The pooling layer is responsible for reducing the feature size of the model for higher computational efficiency [19]. As the number of layers increases, a CNN is considered a Deep CNN. Currently, there are many pre-trained CNN architectures, such as ResNet50, ImageNet, MobileNet, GoogleNet, which have proven to be highly efficient in segmentation and classification. For this reason, in state of the art, many researchers use transfer learning [8] [9] [10] and similar architectures.

Proposed Method

The aim of this research is the classification of skin lesions with a conventional CNN. We have defined two CNN models that can improve the performance of skin lesion classification with the ISIC Archive dataset, compared to the state-of-the-art.

CNN-1. This architecture proposal describes a CNN model of two convolutional blocks. The route begins with the image of the skin lesion, going through two sequential convolutions of 32 features and a 3x3 kernel. The second convolution uses a ReLu activation function [20]. It

is followed by a layer of Max Pooling, whose kernel is 2x2, to reduce the proof of the image entered. This step reduces the dimensionality of the array on which the operations are performed. Also, there is a Dropout Layer with a probability of 0.2 to avoid overfitting [18] when training the model. This process is repeated, with the difference that now the convolutional layers will consist of 64 features and a 3x3 kernel, both with function ReLu activation [20]. Classification is done in a fully connected layer at the end of the path. A Flatten Layer is included in charge of reducing the dimensionality to a vector, and on that, two Dense Layers are used, of 512 and 128 neurons, respectively. Before final classification, it Includes a Dropout Layer with probability 0.2 [18]. The output is binary, from 1 neuron with a Sigmoid activation function.

CNN-2. This proposed architecture describes a CNN model of four convolutional blocks. The route begins with the image of the skin lesion, passing through the two convolutional blocks described in the CNN-1 architecture. This is followed by a third convolutional block consisting of two 128-feature convolutional layers and a 3x3 kernel. In addition, it consists of an Activation Layer with the ReLu function [20], a 2x2 kernel Max-Pooling Layers, and a Dropout Layer with a probability of 0.2 [18]. The fourth convolutional block consists of two convolutional layers of 512 features, a 5x5 kernel, and a ReLu activation function [20]. This is followed by a 4x4 kernel MaxPooling layer and a 0.2 probability Dropout Layer [18]. The fully connected layer of this architecture uses a Flatten Layer to pass the data to a vector. A Dense Layer of 1024 neurons is used, followed by a Dropout Layer of probability 0.2 [18]. Finally, the classification is carried out in a layer of 1 neuron with a sigmoid activation function.

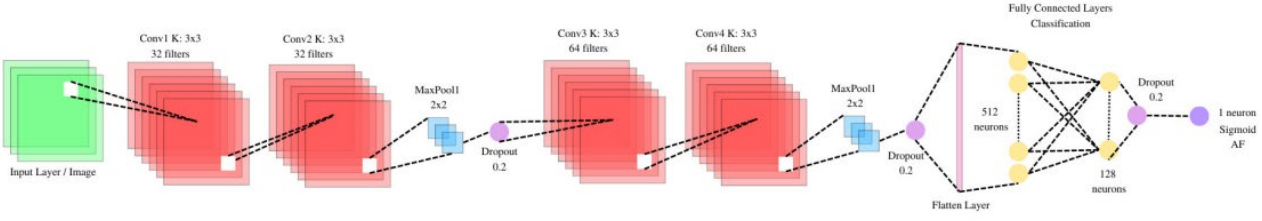

Fig. 2: CNN1 proposed architecture

Experimental Setup

Data processing

The dataset created includes images ranging in size from 354 x 329 to 6669 x 4439 pixels. For this reason, it is required to preprocess the data before feeding it to CNN. The preprocessing consists of resizing the images of the created dataset to a general size of 128 x 128 pixels. In this way, computation time and processing capacity are reduced.

Training, validation and test sets

For training and evaluation, 10% of 2860 images from the experimental dataset were randomly selected, where there are benign skin lesions and melanoma, to form the test set. The remaining 90% is used in training and validation, using the stratified k-cross validation technique, with a k of 10. The use of this technique allows for evaluating the effectiveness of the model against different combinations of data, maintaining the proportion between benign lesions and melanoma, and avoiding overfitting [21]. In this way, we have an overview of how the model performs against the entire data set.

Model Configuration

The hyperparameter configuration consisted of a static batch size of 128 and 500 epochs for both architectures (CNN-1 and CNN-2). The dimension used in training (128 x 128) allows a high batch size to be used. Furthermore, we used a 1x104 learning rate with an ˆ Adam optimizer. According to [22], adaptive optimizers do not require setting a fixed learning rate. Instead, they change according to a history of gradients, so that the learning rate is no longer a hyperparameter to be tuned. According to [22], Adam is the best choice of adaptive optimizers, since it adds the storing of an exponentially decaying average of past gradients similar to momentum.

Assessment metrics

As this is a classification problem, the Receiver Operating Curve - Area Under the Curve (ROCAUC), Accuracy (ACC), Precision (PRE), and Recall (REC) metrics will be used. The main selection metric is ROC-AUC because it represents the ability of the model to distinguish between the two classes [23], in this case, benign lesion or melanoma. The support metrics are ACC, PRE, and REC, in case the models present a similar ROC-AUC. We will compute the mean and standard deviation of the metrics described above over the results, so as to select the best model. A metric that will also be considered will be a loss, as it is a way of indicating how good the model's predictions are at classifying. For this case, binary cross entropy will be used. This metric compares the prediction to the ground truth and penalizes the distance between the output probability and the ground truth [24]. For this reason, it is widely used in binary classification, as is the case.

Selection criterion

We considered checkpoints every 50 epochs for a total of 10 models per architecture (CNN1 CNN-2) during the training stage. The model with the maximum punctuation in ROC-AUC will be selected. We will also consider the maximum punctuation in accuracy if any two models share similar ROC-AUC scores.

RESULTS AND DISCUSSION

According to the experimental configuration designed, the experimental dataset of 2600 training images was evaluated, with the models defined as CNN-1 and CNN-2. The comparison between the ROC-AUC, Binary Cross Entropy Loss, and Accuracy metrics mainly highlights the results for the classification of skin lesions between benign and melanoma, as can be read in Table I.

TABLE I: Performance results of deep learning models.

Architecture	Conv. layer (f)	Kernel size	Pool size per layer	FC layer (n)	Batch size	Epochs (u)	AUC (u)	Loss B.C	ACC (%)	PRE (u)	REC (u)
CNN-1	(32, 32, 64, 64)	(3 × 3)	(2 × 2)	(512, 128, 1)	128	**50**	**0.843 ± 0.05**	**0.574 ± 0.12**	**69 ± 0.6**	**0.76 ± 0.08**	**0.49 ± 0.21**
						100	0.824 ± 0.04	0.790 ± 0.33	68 ± 0.46	0.73 ± 0.11	0.52 ± 0.21
						150	0.841 ± 0.04	1.02 ± 0.45	75 ± 0.44	0.79 ± 0.05	0.62 ± 0.15
						200	0.837 ± 0.05	1.38 ± 0.63	76 ± 0.4	0.79 ± 0.05	0.66 ± 0.11
						250	0.851 ± 0.05	1.55 ± 0.65	79 ± 0.4	0.78 ± 0.04	0.74 ± 0.11
						300	0.851 ± 0.05	1.66 ± 0.63	79 ± 0.4	0.79 ± 0.04	0.75 ± 0.11
						450	0.850 ± 0.04	2.08 ± 0.81	79 ± 0.03	0.77 ± 0.03	0.77 ± 0.09
						500	0.844 ± 0.05	2.26 ± 1.03	80 ± 0.04	0.78 ± 0.03	0.74 ± 0.12
CNN-2	(32, 32, 64, 64, 128, 128, 512, 512)	(3 × 3)(5 × 5)	(3 × 3)	(1024, 1)	128	50	0.855 ± 0.06	0.521 ± 0.10	70 ± 1.0	0.77 ± 0.09	0.47 ± 0.28
						100	0.887 ± 0.04	0.475 ± 0.12	76 ± 0.9	0.79 ± 0.06	0.62 ± 0.22
						150	**0.902 ± 0.03**	**0.531 ± 0.14**	**82 ± 0.4**	**0.82 ± 0.04**	**0.77 ± 0.09**
						200	0.896 ± 0.02	0.767 ± 0.18	82 ± 0.3	0.80 ± 0.03	0.79 ± 0.08
						250	0.895 ± 0.02	0.969 ± 0.27	84 ± 0.3	0.81 ± 0.02	0.84 ± 0.06
						300	0.889 ± 0.02	1.10 ± 0.30	84 ± 0.3	0.82 ± 0.02	0.82 ± 0.08
						350	0.902 ± 0.01	1.01 ± 0.31	85 ± 0.2	0.82 ± 0.01	0.86 ± 0.05
						400	0.895 ± 0.02	1.01 ± 0.19	83 ± 0.2	0.81 ± 0.02	0.83 ± 0.06
						450	0.901 ± 0.02	1.05 ± 0.29	84 ± 0.2	0.83 ± 0.03	0.84 ± 0.06
						500	0.901 ± 0.02	1.07 ± 0.27	84 ± 0.2	0.81 ± 0.01	0.85 ± 0.06

Conv.- convolutional; f- number of filters per layer; n- number of neurons per layer; FC- fully; connected; u- units; B.C- binary cross-entropy; AUC, ACC, PRE, REC, Loss - mean of metrics AUC, accuracy, precision, recall and loss over ten folds.

Performance evaluation in the training set

From Table I, we can see that the AUC, accuracy, precision, and recall scores do not follow a clear trend during the first training epochs. However, these values begin to stabilize around epochs 50 and 100 of both models (CNN-1 and CNN2). In addition, it can be seen that the metrics obtained with the training and validation of the CNN-2 model exceed, from the beginning, the base model used, CNN-1, whose architecture is simpler. In general, the performance of both models is good, taking into account state of the art, without the use of transfer learning. In the case of the CNN-1 model, the best result was obtained at the 50-epoch mark, when an AUC of 0.843 with a standard deviation of 0.05 is maintained.

In the case of the CNN-2 model, the best result was obtained at the 150-epoch mark, when an AUC of 0.902 with a standard deviation of 0.03 is presented. In both cases, the small standard

deviation gives us an idea of the closeness of the values collected over the 10 iterations produced by the use of k-cross fold validation at the time of training.

Also, in Table I, we can see that the scores of the CNN-1 base model grow faster than the proposed CNN-2 method. After 100 epochs, CNN-2 begins to handle more accurate classifications, so the proposed model presents an improvement in the performance of a classifier for the detection of cancerous melanoma. The comparison of both models was made on the validation scores obtained during the training of both cases. However, the model must be tested with a set of unseen data to speak from a generalization and application perspective. Up to this point, the selected model CNN-2 of 150 epochs has shown to have great classification capacity.

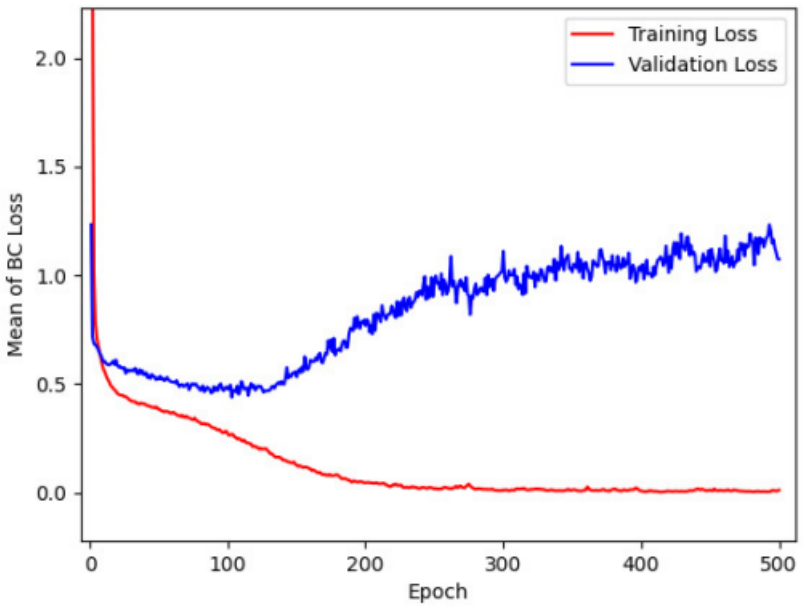

Fig. 3: Mean of Loss vs. Epochs

In Figure 3 we can see the behavior of the best model during its training, taking its loss into account. This graph tells us that the loss values reached a minimum loss score around epoch 150. From that point on, we can see increasing overfitting as more epochs are trained. This may be due to the size of the dataset used to train the model. In the case of having a larger data set, it is likely that overfitting does not occur in that epoch. It is typical of deep learning models

to achieve good training with more than 500 epochs, but it should be noted that these models require a large amount of data to work with to achieve such results. Figure 4, on the left, shows the mean ROC-AUC curve obtained from the validation of the selected model. From the 10 training iterations, an average AUC of 0.88 with a standard deviation of 0.04 was obtained, which proves to be a robust model that performs well on different data sets. Figure 4, on the right, shows the trade-off between precision and recall in the model's training. The relation between these two metrics shows how the model performs predicting correct values. As we can see in Table I, the selected model scores 0.82 and 0.77 for precision and recall, respectively, meaning that this model performs well on classifying between benign skin lesions and melanoma.

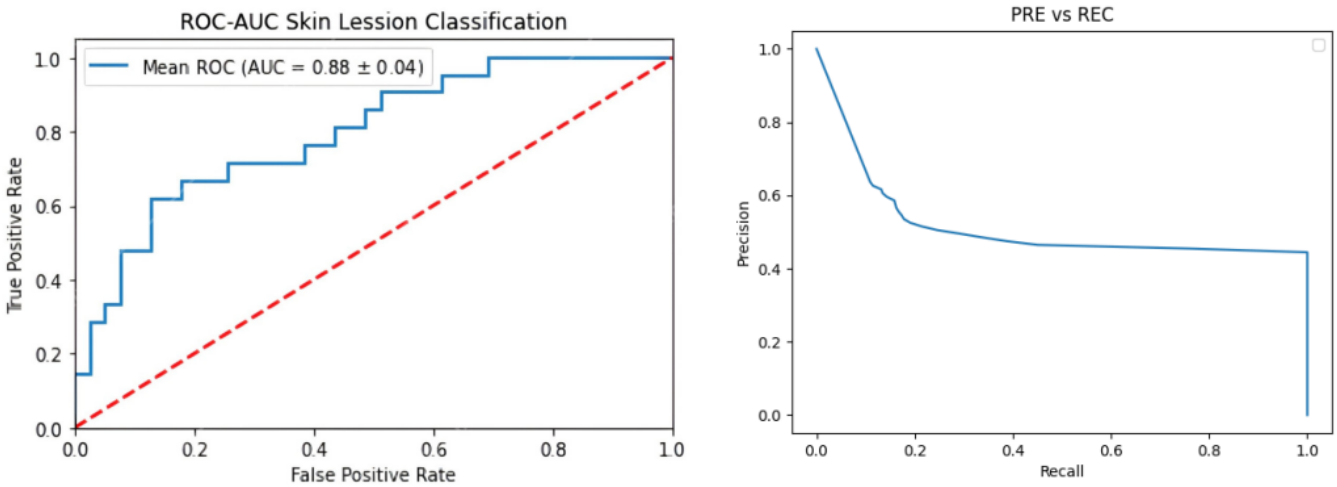

Fig. 4: To the left, ROC-AUC Curve. To the right, Precision vs. Recall Curve

Performance evaluation in the test set

The best model was obtained, CNN-2 with 150 epochs, selected in the training phase based on its scores in the described metrics. To test its generalization power of it, the best model was evaluated on a test set containing images of skin lesions classified as benign and melanoma in equal proportions. The results obtained with this data set are promising. An AUC of 0.9604, an accuracy of 91.54%, a precision of 0.8971, and a recall of 0.9385 was obtained. Figure 5 shows

various results of the classification of skin lesions with the selected model, with the first four columns being correct classifications and the last column being an incorrect classification. Columns 1 and 2 were correctly classified as benign skin lesions. Columns 3 and 4 were correctly classified as melanoma. Column 5 represents a wrongful classification, the lesion is benign, and it was classified as melanoma.

Despite the good results, there are certain skin lesions that remain a challenge for the disorder model. This may be due to changing dimensions of the images in the original data set. As it is made up of large images, resizing to a small size, such as 128x128 pixels, translates into a loss of important information, such as the dispersion of skin pigmentation, which is characteristic and necessary to determine whether a lesion is a cancerous melanoma. For instance, in Figure 5, a small lesion was wrongfully classified as melanoma. In general, the results obtained show that the model is competitive with the options shown in state of the art, and they have a great capacity for generalization.

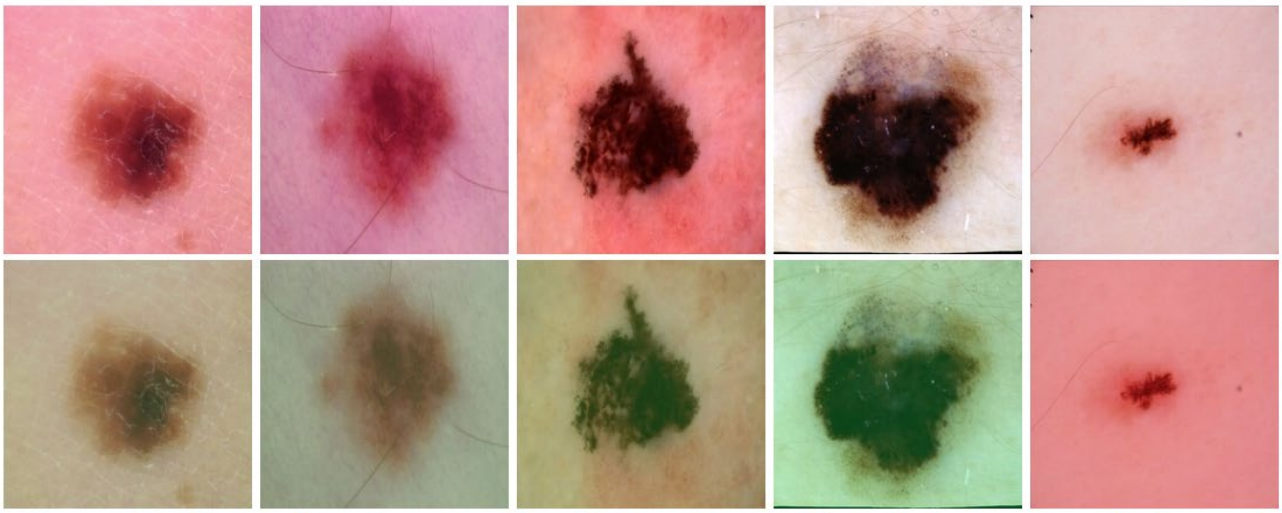

Fig. 5: Some samples of the test classification results: columns 1 & 2 were benign and classified as benign, columns 3 & 4 were melanoma and classified as melanoma, and column 5 was benign and classified as melanoma. Green represents and accurate classification. Red represents a wrong classification.

CONCLUSIONS AND FUTURE WORK

We analyzed two different models for the classification of skin lesions. Both had promising results, but we selected the CNN-2 model, given its scores on the ROC-AUC, ACC, PRE, and REC metrics. Both models were trained under the same conditions, using a stratified 10 cross-fold validation scheme on a dataset of more than 2600 images containing benign skin lesions and melanomas. We analyzed the AUC scores in both models, 0.843 ± 0.05 for CNN-1, and 0.902 ± 0.03 for CNN-2. This demonstrated that the CNN-2 model tends to perform better than the selected reference method, CNN-1. Furthermore, the CNN-2 model with 150 training epochs was validated on an external test set, reaching scores of 0.9604 for AUC. This external performance evaluation confirms the generalization quality of the model, so it can be stated that the model tends to be successful in classifying different types of skin lesions.

In the future, we aim to implement new variations of CNN models for the classification of skin lesions, testing with new architectures or different training parameters. Furthermore, we will explore the combination of this model with transfer learning to improve the performance of the proposed method. Finally, it is necessary to find more extensive databases containing dermoscopic images of melanoma, taking into account that it is a relatively rare skin lesion, in order to achieve better training and thus find a more generalizable model. This process is known as data quality checking [25] and must be done to have better results, even with a simpler CNN. Another plausible approach is to apply data augmentation, which is defined as the process of creating or modifying synthetic data using real data [25], to have a more reliable CNN able to distinguish between these two classes.

ACKNOWLEDGMENT

The authors thank to the Applied Signal Processing and Machine Learning Research Group of USFQ for providing the computing infrastructure (NVidia DGX workstation) to implement and execute the developed source code. The publication of this article was funded by the Academic Articles Publication Fund of Universidad San Francisco de Quito USFQ.

REFERENCES

[1] B. C. Ortega, Dermatoscopia: utilidad y peculiaridades en piel pigmentada. McGraw Hill Interamericana, 2016.

[2] A. E. Acosta, E. Fierro, V. E. Vel ´ asquez, and X. Rueda, "Melanoma: ´ patogenesis, cl ´ ´ınica e histopatolog´ıa," Revista de la Asociacion´ Colombiana de Dermatolog´ıa y Cirug´ıa Dermatologica ´ , vol. 17, no. 2, pp. 87–108, 2009.

[3] F. Solca, "Registro nacional de tumores." [Online]. Available: https://solcaquito.org.ec/registro-nacional-de-tumores/#: „:text=El%20Registro%20Nacional%20de%20Tumores,en%20la% 20ciudad%20de%20Quito.

[4] S. Nu´nez-Gonz ˜ alez, E. Bedoya, D. Simancas-Racines, and C. Gault, ´ "Spatial clusters and temporal trends of malignant melanoma mortality in ecuador," SAGE Open Medicine, vol. 8, p. 2050312120918285, 2020.

[5] M. d. l. A. M. O ´ nate, M. A. R. Pinto, A. G. A. Galarza, and N. V. P. ˜ Mena, "Melanoma cutaneo," ´ RECIMUNDO, vol. 6, no. 4, pp. 77–86, 2022.

[6] M. A. Arasi, E.-S. M. El-Horbaty, A. El-Sayed et al., "Classification of dermoscopy images using naive bayesian and decision tree techniques," in 2018 1st Annual International Conference on Information and Sciences (AiCIS). IEEE, 2018, pp. 7–12.

[7] J. Almaraz-Damian, V. Ponomaryov, and E. Rendon-Gonzalez, "Melanoma cade based on abcd rule and haralick texture features," in 2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW). IEEE, 2016, pp. 1–4.

[8] A. Sagar and D. Jacob, "Convolutional neural networks for classifying melanoma images," bioRxiv, pp. 2020–05, 2021.

[9] A. Sallam, A. E. Ba Alawi, and A. Y. Saeed, “A cnn-based model for early melanoma detection,” in International Conference of Reliable Information and Communication Technology. Springer, 2020, pp. 41– 51.

[10] M. F. Jojoa Acosta, L. Y. Caballero Tovar, M. B. Garcia-Zapirain, and W. S. Percybrooks, “Melanoma diagnosis using deep learning techniques on dermatoscopic images,” BMC Medical Imaging, vol. 21, no. 1, pp. 1–11, 2021.

[11] J. Song, J. Li, S. Ma, J. Tang, and F. Guo, “Melanoma classification in dermoscopy images via ensemble learning on deep neural network,” in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020, pp. 751–756.

[12] J. Xingguang, W. Yuan, Z. Luo, Z. Yu et al., “Deep neural network for melanoma classification in dermoscopic images,” in 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE, 2021, pp. 666–669.

[13] T.-C. Pham, C.-M. Luong, M. Visani, and V.-D. Hoang, “Deep cnn and data augmentation for skin lesion classification,” in Asian Conference on Intelligent Information and Database Systems. Springer, 2018, pp. 573–582.

[14] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. R. Soroushmehr, M. H. Jafari, K. Ward, and K. Najarian, “Melanoma detection by analysis of clinical images using convolutional neural network,” in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016, pp. 1373–1376.

[15] A. A. Adegun and S. Viriri, “Deep learning-based system for automatic melanoma detection,” IEEE Access, vol. 8, pp. 7160–7172, 2019.

[16] [17] F. Chollet, Deep learning with Python. Simon and Schuster, 2021.

[18] J. Moolayil, Learn Keras for Deep Neural Networks. Apress, 2019.

[19] S. Raschka and V. Mirjalili, “Python machine learning: Machine learning and deep learning with python,” Scikit-Learn, and TensorFlow. Second edition ed, vol. 3, 2017.

[20] S. Sharma, “Activation functions in neural networks,” Nov 2022. [Online]. Available: https://towardsdatascience. com/activation-functions-neural-networks-1cbd9f8d91d6#:,,: text=The%20main%20reason%20why%20we,sigmoid%20is%20the% 20right%20choice.

[21] K. Muralidhar, “What is stratified cross-validation in machine learning?” Feb 2021. [Online]. Available: https://towardsdatascience.com/ what-is-stratified-cross-validation-in-machine-learning-8844f3e7ae8e

[22] D. Giordano, “7 tips to choose the best optimizer - towards data science.” [Online]. Available: https://towardsdatascience.com/ 7-tips-to-choose-the-best-optimizer-47bb9c1219e

[23] S. Narkhede, “Understanding auc - roc curve - towards data science.” [Online]. Available: https://towardsdatascience.com/ understanding-auc-roc-curve-68b2303cc9c5

[24] D. Godoy, “Understanding binary cross-entropy / log loss: A visual explanation ...” [Online]. Available: https://towardsdatascience.com/ understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181

[25] ISO, “ISO/IEC 22989:2022(en) Information technology — Artificial intelligence - Artificial intelligence concepts and terminology” [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en

ANEXO A: SOURCE CODE - GITHUB

El código fuente de la investigación presentada se encuentra en el siguiente repositorio:

https://github.com/EvilJKD/melanoma-classification