Skin Lesion Analysis for Melanoma Detection using Neural Networks

Dwight Louis H. Velasco ,*[1] Misha Hilario,[1] Kathleen Mae V. Edquila,[1] and Mikamila Elehn-joyce Garcia[1]

[1] *Department of Physical Sciences and Mathematics, College of Arts and Sciences, University of the Philippines Manila*
*Corresponding author: dhvelasco@up.edu.ph

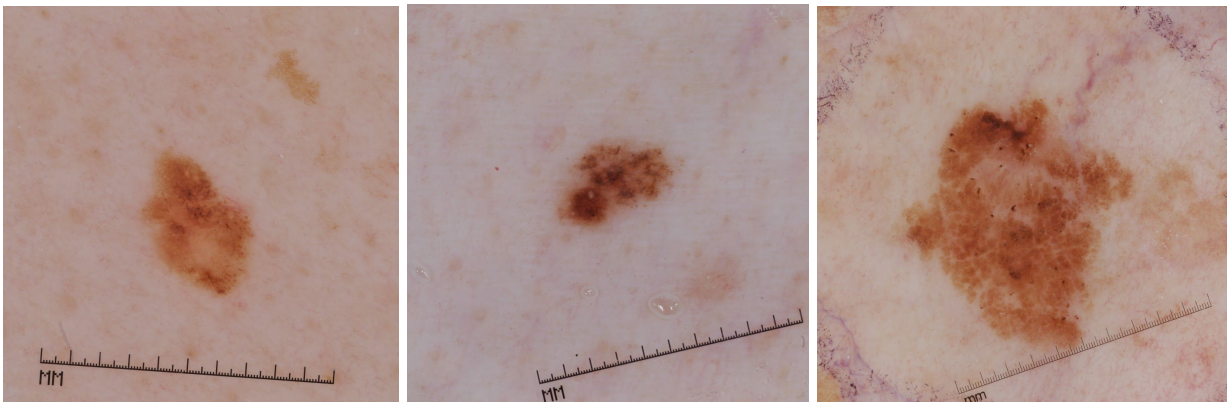
Abstract

Melanoma is a type of skin cancer that develops by mutation of pigment cells known as melanocytes. The survival rate for melanoma is dependent on the cancer cell mutation. Melanoma is highly curable when detected early. In this project, we proposed a deep learning system for melanoma detection. This project aims to diagnose melanoma and differentiate malignant skin tumor from two types of benign lesions: (1) Benign nevus, and (2) Seborrheic keratosis. Two tasks were assigned to the training model, task 1 is to differentiate melanoma from benign skin lesions, and task 2 is to differentiate Nevi from Seborrheic keratoses. Data from results show that an accuracy rate of 78% was obtained for task 1 and 91% accuracy rate for task 2. The model can be further developed for possible implementation in clinical diagnosis.

Keywords: melanoma, deep learning, convolutional neural networks

1 Introduction

Skin cancer is a major health problem in the Philippines, especially in rural regions which lack access to health services provided by experienced dermatologists. Melanoma is one the deadliest form of skin cancer, which causes a tumour in the melanin-forming cells [1]. Melanoma is less prevalent in Philippines than it is in other countries, but the rate at which people are getting affected is increasing. Early detection is critical, as the estimated 5-year survival rate for melanoma drops from over 99% if detected in its earliest stages to about 14% if detected in its latest stages [2].



Figures 1, 2, 3. (L-R): Melanoma, Nevus, Seborrheic Keratosis.

Melanoma is a type of malignant skin cancer while Nevus and Seborrheic Keratosis are types of benign skin lesions. Skin lesions are usually similar in appearance. As observed in Figures 1, 2, and 3, there is a similar appearance between different types of skin lesions characterized by brownish pigments.

The resemblance of these images makes it difficult to differentiate malignant and benign lesions. This may result to misdiagnosis of the patient which can cause possible complications. In response to this problem, this project proposes to diagnose melanoma and differentiate melanoma from the two types of benign skin lesion, Nevus and Seborrheic Keratosis.

We used a convolutional neural network (CNN) for skin lesion detection. A convolutional neural network (CNN) is one of the most popular algorithms for deep learning, a type of machine learning in which a model learns to perform classification tasks directly from images, video, text, or sound [3].

CNNs are particularly useful for finding patterns in images to recognize objects, faces, and scenes. They learn directly from image data, using patterns to classify images and eliminating the need for manual feature extraction. Applications that call for object recognition and computer vision — such as self-driving vehicles and face-recognition applications — rely heavily on CNNs [3].

CNNs provide an optimal architecture for image recognition and pattern detection. For example, deep learning applications use CNNs to examine thousands of pathology reports to visually detect cancer cells. By creating a disease-partitioning algorithm that maps individual diseases into training classes, we are able to build a deep learning system for automated dermatology .

2 Methodology

Convolutional Neural Network (CNN) Architecture

This section discuss the details of the CNN. CNN is a deep learning framework which was used for automatic detection of melanoma [4]. They can examine various structures in input images however, in utilization of CNN, the input is the image itself and the network automatically extracts appropriate aspects of the image.

However, extracting unique feature set is challenging since it can lead to feeding some incoherent traits to the network or the possibility of missing some proper descriptors. Nonetheless, utilization of this automatic feature extraction systems can achieve a discriminative feature set based on labeled training set without the need for definition of handcrafted feature extraction procedures.

The CNN, like other neural networks, is composed of an input layer, an output layer, and many hidden layers in between. These layers perform operations that alter the data with the intent of learning features specific to the data. Three of the most common layers are: convolution, activation or ReLU, and pooling [3].

Convolution puts the input images through a set of convolutional filters, each of which activates certain features from the images. Rectified linear unit (ReLU) allows for faster and more effective training by mapping negative values to zero and maintaining positive values. This is sometimes referred to as activation, because only the activated features are carried forward into the next layer. Pooling simplifies the output by performing nonlinear downsampling, reducing the number of parameters that the network needs to learn. Passing an image through a series of these operations outputs a feature vector containing the probabilities for each class label. Note that in this setup, we categorize an image as a whole. That is, we assign a single label to an entire image. After learning features in many layers, the architecture of a CNN shifts to classification. The next-to-last layer is a fully connected layer that outputs a vector of K dimensions where K is the number of classes that the network will be able to predict. In our case, K = 3. This vector contains the probabilities for each class of any image being classified. The final layer of the CNN architecture uses a classification layer to provide the classification output [3].

CNNs usually use a larger number of datasets for properly training, however, this study uses a limited number of images for detection of melanoma from non-dermoscopic images.

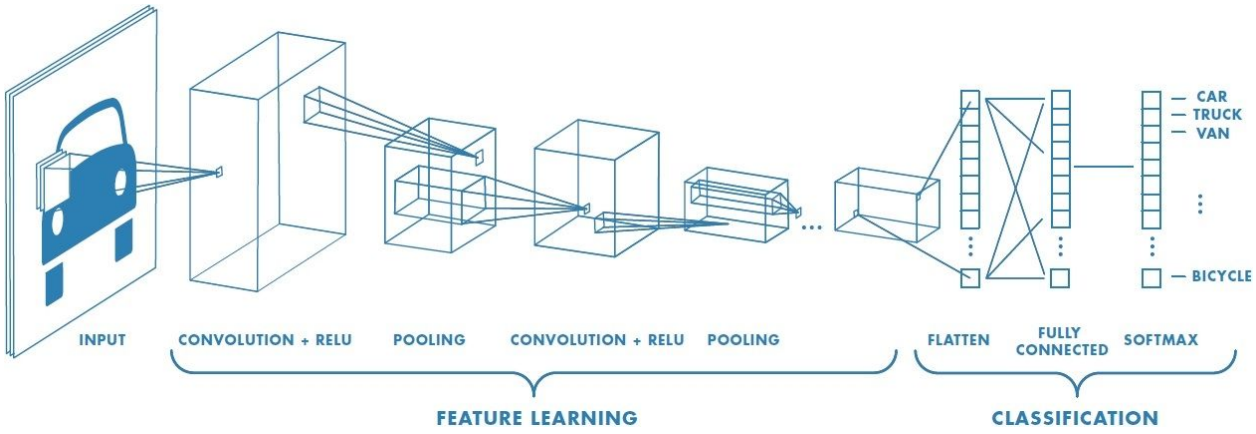


Figure 4. Generalized deep convolutional neural network architecture. Filters are applied to each training image at different resolutions, and the output of each convolved image is used as the input to the next layer [3].

Sample Images

This study used 2000 (374 melanoma, 254 seborrheic keratosis, and 1372 benign nevi) training images, 150 (30 melanoma, 42 seborrheic keratosis, and 78 benign nevi) validation images and 600 (117 melanoma, 90 seborrheic keratosis, and 393 benign nevi) testing images. The total number of images used is based on the availability of data from the following address: http://challenge2017.isic-archive.com/

Transfer Learning

To reduce training time without sacrificing accuracy, we train the CNN using Transfer Learning — which is a method that allows us to use neural networks that have been pre-trained on a large dataset. This approach is commonly used for object detection, image recognition, speech recognition, and other applications [5]. As seen in Figure 5, by keeping the early layers and only training newly added layers, we are able to tap into the knowledge gained by the pre-trained algorithm and use it for our application.

The keras package (v.2.2.0) includes several pre-trained deep learning models that can be used for prediction, feature extraction, and fine-tuning.

The CNN was built on keras with a tensorflow-gpu (v.1.9.0) backend. The transfer learning model is the Resnet50 model pre-trained on ImageNet, which, trained on 14+ million images, can classify an object from one of 22,000+ categories.

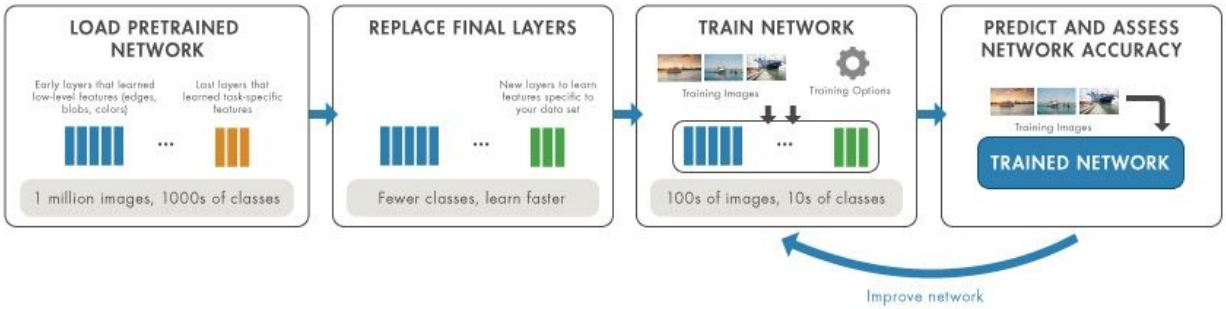


Figure 5. Generalized transfer learning workflow [5].

After modifying the ResNet50 model by replacing the default fully-connected top layers with a Global Average Pooling layer and 4 fully-connected layers, the model ends up with a total of 5,509,635 trainable parameters. The last dense layer contains one node for each skin lesion category (i.e. 3 nodes). Hyperparameter tuning was then done by setting different configurations on the number of nodes in the first, second, and third dense layers and the optimizer learning rate by looking at the loss curve. Given an image, this modified ResNet50 model returns a prediction for the skin lesion that is contained in the image.

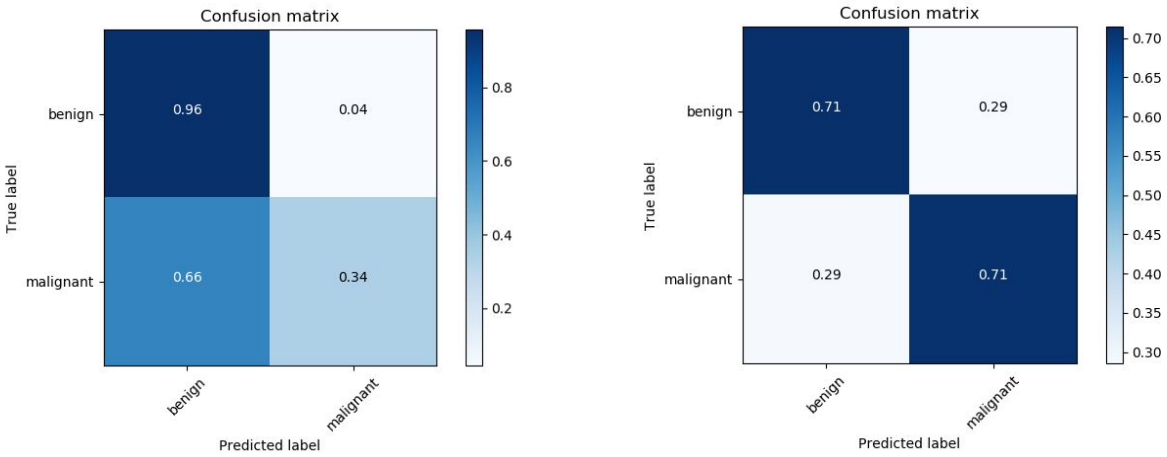
Disease Classification

The sample images were fed to the CNN. The images were classified into 3 categories: "melanoma", "seborrheic keratosis", and "benign nevus", with classification scores normalized between 0.0 to 1.0 for each category (and 0.5 as the binary decision threshold). The model iterated for 43 epochs in which the model's hyper-parameters were fine-tuned via the optimizer in order to reduce the loss function. Mean values of results were then reported.

3 Results and Discussion

A confusion matrix is a table that is often used to summarize and describe the performance of a classification model [6]. It was observed, based from the obtained confusion matrix shown in Figure 1, that benign moles were better detected with sensitivity = $\frac{True\ Benign}{Actual\ Benign} = \frac{0.95}{1} = 0.95$, greater than compared to malignant moles with sensitivity = $\frac{True\ Malignant}{total} = \frac{0.34}{1} = 0.34$. The accuracy of the classifier, at threshold = 0.5, was calculated using accuracy binary classification: $\frac{True\ Benign + True\ Malignant}{total} = \frac{0.95 + 0.34}{2} = 0.65$, obtaining 65 % accuracy. The low accuracy rate is due to the high detection error for malignant skin lesions.

In Figure 2, the accuracy of the classifier, at threshold = 0.3, was calculated using accuracy binary classification: $= \frac{0.71 + 0.71}{2} = 0.71$, obtaining 71 % accuracy. The accuracy improved as threshold was lowered due to the fact the average task 1 score for melanoma images was 0.41.



(a) (b)

Figure 2. Two-by-two Confusion Matrix (a) with threshold = 0.5 and (b) with threshold = 0.3.

Figure 3 shows a 3x3 confusion matrix for three different classes of skin lesions namely, melanoma, nevi, and seborrheic keratoses. The classifier made a total of 600 predictions in accordance with the 600 available images used for testing. The accuracy of the resulting confusion matrix was calculated using formula: $\frac{True\ Melanoma + True\ Nevus + True\ Seborrheic\ keratosis}{total\ predictions\ or\ testing\ images} = \frac{55+314+66}{600} = 0.725$, obtaining an accuracy rate of 72.5 %. Among the three classes of skin lesions, the nevus class has the the highest sensitivity with 0.80, calculated using the sensitivity formula: $\frac{True\ Nevus}{Actual\ Nevus} = \frac{314}{393} = 0.80$, followed by seborrheic keratoses with 0.73, and lastly melanoma with 0.47 sensitivity. This means the probability of correctly identifying the nevus skin lesions, using the proposed classification model, is higher compared to detecting seborrheic keratoses and melanoma.

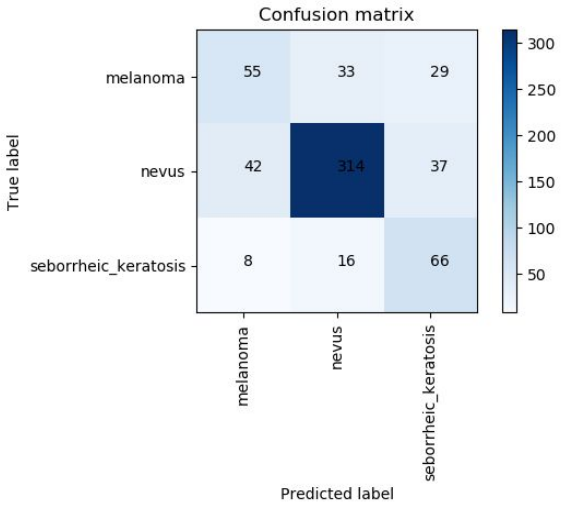


Figure 3. Three-by-three confusion matrix.

The proposed classification model was initially trained on a training dataset composed of images of different types of skin lesions to adjust the parameters required for model fitting. A second dataset, the validation dataset, was used on the fitted model to check for overfitting and to compare the performances of various algorithms and choose the best performing one before conducting the final testing of the model.

Training accuracy is the accuracy of a model on examples that it was constructed on, while validation accuracy is the number of data points classified correctly when the model was applied on the validation data [7]. Figure 4 shows the classification model's accuracy when applied to the training data and validation data. In contrast to accuracy, training and validation loss shown in Figure 5, is the summation of errors made for each image example in the training and validation dataset. It can be observed that the increase in accuracy corresponds to a decrease of loss in training and validation. The graphs shown in figure 4 and 5 imply a well performing model since the accuracy for both pre-testing datasets were increasing.

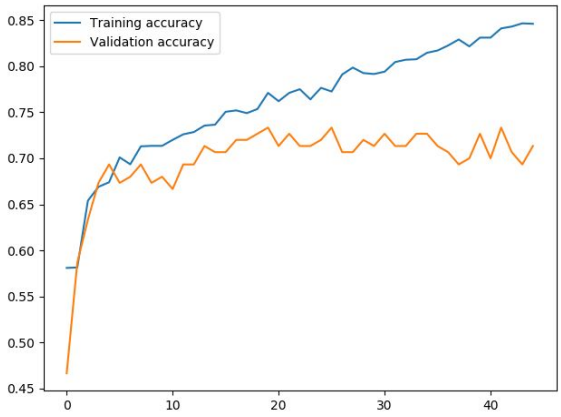


Figure 4. Pre-testing accuracy of the model in terms of training (blue) and validation (orange) accuracy.

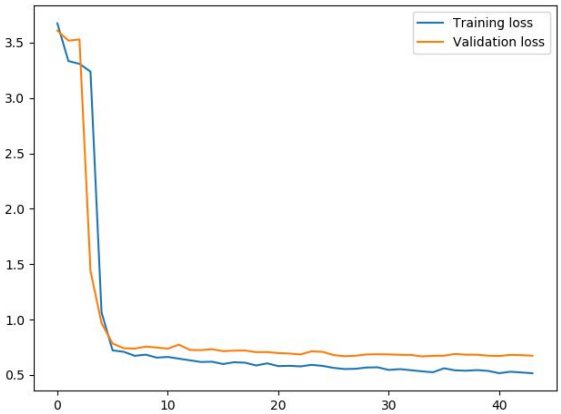


Figure 5. Pre-testing loss of the model in terms of training (blue) and validation (orange) loss.

The diagnostic performance of the classification model, or the accuracy of a test to discriminate malignant cases from benign cases is evaluated using Receiver Operating Characteristic (ROC) curve analysis. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between diagnostic groups [8]. Based from the ROC curve in the discrimination of the three classes of skin lesions shown in Figure 6, the curve for seborrheic keratosis is the closest to the upper left corner, indicating that the classification model for the detection of seborrheic keratoses is more accurate than for the nevus and melanoma. The number of testing images for seborrheic keratosis, which is lesser compared to the other two skin lesions, may have been one of the factors that led to an easier classification of seborrheic keratosis in the dataset, resulting to a higher accuracy rate for the classification of seborrheic keratosis.

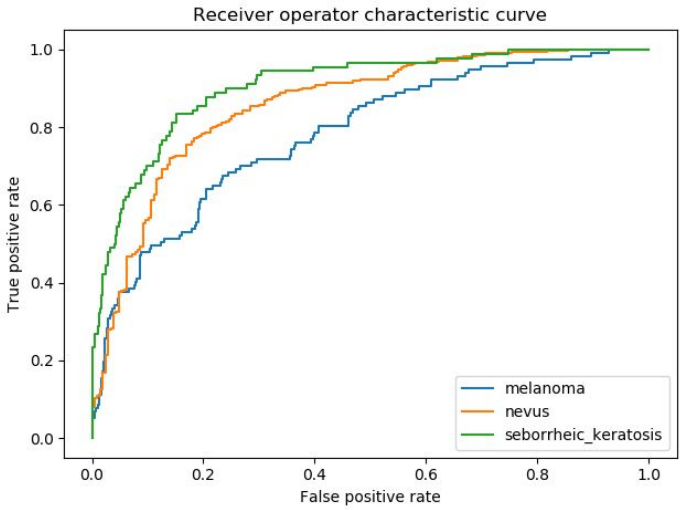


Figure 6. ROC curves for melanoma (blue), nevus (orange), and seborrheic keratosis (green).

Figure 7 shows a comparison of two ROC curves designated as *task 1* and *task 2*. Task 1 is the model's predicted probability that the image depicts melanoma over benign skin lesions, while Task 2 is the model's predicted probability that the image depicts seborrheic keratosis compared to nevus. Based on the figure, the AUC of *task 2* (AUC=0.91) is greater than the AUC of task 1 (AUC=0.78), implying that the test for detecting seborrheic keratosis over nevus class has a more accurate diagnostic performance and thus is a more useful model than the test for detecting melanoma over benign skin lesions.

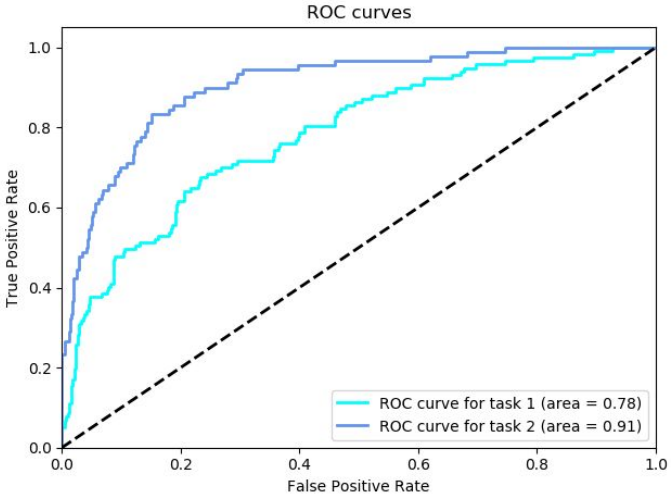


Figure 7. ROC curves for *task 1* (melanoma vs non-cancerous lesions) and *task 2* (seborrheic keratosis vs nevus class).

4 Conclusion

The model for automated dermatology was created and trained using convolutional neural networks. Two tasks were assigned to the model: (1) detection of melanoma from benign skin lesions and (2) detection of Nevus from seborrheic keratosis. Current findings show that an accuracy rate of 91% was obtained in task 2 which is greater than the accuracy rate of 78% obtained in task 1. This indicates that the model has a better performance in detecting nevus over seborrheic keratosis than the detection of melanoma over benign skin lesions. The number of testing images can be a possible factor of accuracy rate for each task. The number of testing images used for melanoma (374) is less than the number of testing images for benign lesions (254 seborrheic keratoses and 1372 benign nevi). The deep learning model can be tested and further developed with a larger dataset to improve its accuracy and performance.

References

[1] The Skin Cancer Foundation, Melanoma, 2018 Retrieved from https://www.skincancer.org/skin-cancer-information/melanoma

[2] F..Lowry, "AI as Good as Docs for Diagnosing Skin Cancers From Photo", Medscape, 2017 https://www.medscape.com/viewarticle/87494

[3] Convolutional Neural Network, Deep Learning, Retrieved from https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html

[4] 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, 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, 2016, pp. 1373-1376. doi: 10.1109/EMBC.2016.7590963

[5] Transfer Learning - MATLAB & Simulink. Retrieved from https://www.mathworks.com/discovery/transfer-learning.html

[6] Simple guide to confusion matrix terminology. Retrieved from https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[7] About Train, Validation and Test Sets in Machine Learning. Retrieved from https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

[8] ROC curve analysis. Retrieved from https://www.medcalc.org/manual/roc-curves.php