Pre-Processing – ISIC 2019

# Overview

ISIC 2019 dataset contains images of HAM10000 and the BCN\_20000. HAM10000 contains 10000 images with a size of **600 x 450**. This dataset was the older challenge of ISIC 2018. While the BCN\_20000 contains 19424 images of size **1024 x 1024**.

# Data Description

The ISIC – 2019 training dataset consists of 25331 dermatoscopic images of skin lesions belonging to 8 different cases of skin cancers. The ground truth of these images is stored in another dataset which consists mapping of these images to their corresponding labels (skin lesion class). There are 8 cases in the 2019 namely:

* MEL - Melanoma
* NV - Melanocytic Nevus
* BCC -  Basal Cell Carcinoma
* AK -  Actinic Keratosis (referred as AKIEC in 2018)
* BKL - Benign Keratosis Lesion
* DF - Dermatofibroma
* VASC - Vascular Lesions
* SCC - Squamous Cell Carcinoma

# Pre- Processing Steps

The goal of the pre-processing is to improve the training data for CNN model by removing undesirable noise and enhancing the images to improve the performance of the model. The following steps were taken in the pre-processing stage:

1. Mapping – Organize images for efficient importing
2. Hair Removal – Eliminate element of noise that might be learned as a feature by model
3. Resizing – Reduces time taken at each epoch of training CNN model.
4. Cropping – Avoid unwanted areas of image (dark corners, ruler marks, normal skin area)
5. Augmentation – Balance the amount of data in each class and increase the training data.
6. Normalization – Ensure each pixel has similar distribution making convergence of CNN faster.

## Step 1: Mapping

With the help of python script all the 25331 training set images from the 2019 dataset which were contained in a single folder were moved to their respective directories according to the mapping present in the ground-truth dataset. The counts of images in each class after mapping are shown below:

|  |  |
| --- | --- |
| Class | COUNT |
| NV | 12875 |
| MEL | 4522 |
| BKL | 2624 |
| BCC | 3323 |
| AKIEC | 867 |
| VASC | 253 |
| DF | 239 |
| SCC | 628 |
| Total | 25331 |

## Step 2 : Digital Hair Removal

Artifacts such as hair especially dark coloured hair are the major obstacles for creating a robust model. A CNN model can learn every feature of image given to it as training data. Hair in images is processed as a feature of skin lesion itself resulting in poor model performance. Hence, it becomes essential to eliminate any such noise.

The following methodology has been employed to tackle this problem:

* A different approach was formulated for removing hairs from the 2019 dataset. As the size of images in each class is more than twice the number of images present in the 2018 dataset, it was decided that this pre-processing step will be applied class wise. We visually inspected all 8 classes to check where this problem exists predominantly. It was found that the following 5 classes had most images with hairs.
  + MEL – Melanoma
  + BKL - Benign Keratosis Lesion
  + NV - Melanocytic Nevus
  + BCC - Basal Cell Carcinoma
  + VASC - Vascular Lesions
* The techniques used for hair removal is called Dullrazor which is a well-known mechanism for removing hair from dermatoscopic images. The algorithm has been developed after research and is efficient especially for skin cancer images. It performs better on especially dark thick hair as they are the main culprit that confuses the CNN model resulting in unsatisfactory results. The Dullrazor[[1]](#footnote-1) algorithm and its working are described briefly :
  + The hair regions are initially detected through the morphological closing operator on each RGB color channel separately and with three structuring elements having different directions.
  + To generate the binary mask, a thresholding process is applied to the absolute difference between the original color channel and the image generated by the closing .
  + The mask pixels undergo a bilinear interpolation between two nearby not-mask pixels.
  + Finally, to the resulting image, an adaptive median filter is applied.

We have implemented this algorithm in Python using OPENCV library.

* Results after applying the hair removal pre-processing is shown below:

A picture containing turtle

Description automatically generated**MEL:**

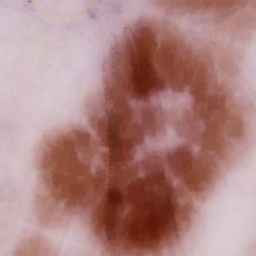
A picture containing snow, close

Description automatically generated





**BKL:**





**NV:**



A close up of a person's eye

Description automatically generated with medium confidenceA picture containing star, weapon, outdoor object, blurry

Description automatically generated

**BCC:**



A picture containing arthropod, invertebrate

Description automatically generatedA picture containing weapon, blurry, missile

Description automatically generatedA close up of a person's skin

Description automatically generated with low confidence

**VASC:**

****

A picture containing close, ocean floor

Description automatically generatedA picture containing worm, arthropod, acarine

Description automatically generatedA close up of a person's eye

Description automatically generated with medium confidence

**Note** : Dull razor algorithm does not work efficiently with white or light-coloured hair. But they are automatically handled by the CNN and are not as much a hinderance as compared to dark thick hair.

## Step 3 : Resizing

Resizing images is a critical step in Image pre-processing as CNN models run faster on smaller images. In the 2019 dataset, 10000 images have size of 600 x 450, and the remaining 19424 images of size 1024x1024. The resizing has been applied during the augmentation process in the 2019 dataset and the resulting images are stored with a resolution of 256 x 256. The resulting images after augmentations have a size of about 150 MB per class. We consider square dimensions as the Tensor Flow library which we are using to create our model, expects that dimension.

## Step 4 : Image Cropping

The image cropping for 2019 dataset images was applied to only 4 classes as a last augmentation step. For classes AK and SCC, cropping with size 175 x 175, was applied to only a selected number of images which were shuffled and randomly selected. For classes DF and VASC, the centre crop augmentation was performed at the beginning with a crop size of 125 x 125. We chose a smaller crop size for these two classes as the lesions here are mostly small in shape than compared to the other two classes mentioned above. Likewise, a bigger resolution of 175 x 175 was chosen to accommodate a slightly bigger lesion size of the classes AK and SCC.

The cropping was not applied to remaining classes because the lesions in the remaining classes were uneven shaped and spread throughout the dimension of the image. Therefore, cropping such images would have resulted in elimination of important features for the model

Overview of Image Cropping Results:

**AK**:

A close up of a planet

Description automatically generated with low confidenceA close up of a person's skin

Description automatically generated with low confidence

SCC:

A close up of a person's face

Description automatically generated with low confidenceA close up of a person's face

Description automatically generated with low confidence

DF:

A close up of a person's eye

Description automatically generated with medium confidenceA close up of a person's face

Description automatically generated with low confidence

VASC:

A close up of a planet

Description automatically generated with low confidenceA close up of a person's eye

Description automatically generated with medium confidence

## Step 5 : Augmentation

Augmentation of images is done to deal with the problem of skewed classes , overfitting, and training image scarcity. For 2019 dataset, augmentations are applied to match the training size of the biggest class NV, size = 12875. We shall consider a training input of 70% = 0.7x12875 = 9012 for this class. Therefore, for all other classes, necessary and suitable augmentations are performed to match a range of (8700, 9300). This process was performed with a focus that the resulting images are obtained as natural, real, as possible and avoid duplications.

The details, sequence and parameters of augmentations for the 2019 dataset are summarized in the file ” **ISIC\_2019\_augmentations\_summary.xlsx** ” and all relevant codes are uploaded in the Github Repository. Kindly refer the file mentioned above for the augmentation process.

***Glimpse of the Augmentation Results***

**AK**:

A close up of a person's skin

Description automatically generated with low confidenceA close up of a person's skin

Description automatically generated with low confidenceA close-up of a person's skin

Description automatically generated with low confidence A close up of a person's foot

Description automatically generated with low confidenceA close-up of a person's skin

Description automatically generated with low confidenceA close up of a person's skin

Description automatically generated with low confidence

**DF**:

A close up of a person's skin

Description automatically generated with low confidenceA close up of a person's face

Description automatically generated with medium confidence A close up of a person's face

Description automatically generated with low confidenceA close up of a person's face

Description automatically generated with low confidenceA close up of a person's skin

Description automatically generated with low confidence

**VASC**:

A close up of a person's skin

Description automatically generated with low confidence

**BCC**:

A close up of a person's skin

Description automatically generated with low confidenceA close up of a person's skin

Description automatically generated with low confidence

**BKL**:

A close up of a person's face

Description automatically generated with medium confidenceA picture containing person, blurry

Description automatically generatedA close up of a person's face

Description automatically generated with medium confidence

**MEL** :

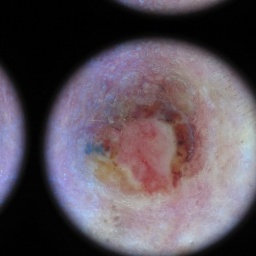
A picture containing indoor

Description automatically generatedA picture containing indoor

Description automatically generated

**SCC** :

A picture containing fig

Description automatically generatedA close-up of a person's chest

Description automatically generated with low confidenceA close up of a planet

Description automatically generated with low confidence

## Step 6 : Normalization

The final step of pre-processing is standardizing each pixel value of image from 0-255 to 0-1 called Normalization. The main objective of performing normalization is to achieve consistency in dynamic range of pixel intensity of images of different skin lesions. Furthermore, it will help CNN model converge faster. Normalization will be applied while developing the model as it comes as a built-in feature of ImageDataGenerator class in Tensorflow library of python.

1. Lee, T.; Ng, V.; Gallagher, R.; Coldman, A.; McLean, D. Dullrazor: A software approach to hair removal from images. *Comput. Biol. Med.* **1997**, *27*, 533–543. [↑](#footnote-ref-1)