Deep Learning and Transfer Learning for Skin Cancer Segmentation and Classification

Lin Li
Department of Computer Science
Seattle University
Seattle, USA
lil@seattleu.edu

Wonseok Seo
Department of Computer Science
Seattle University
Seattle, USA
seow@seattleu.edu

2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE) | 978-1-6654-4261-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBE52308.2021.9635175

***Abstract*—According to Skin Cancer Foundation, skin cancer is by far the most common type of cancer in the United States and worldwide. Early diagnosis of skin cancer is critical because proper treatment at early stages can increase the chance of cure and recovery. However, visual inspection of dermoscopic images by dermatologists is error-prone and time-consuming. To ensure accurate diagnosis and faster treatment of skin cancer, deep learning techniques have been utilized to conduct automated skin lesion segmentation and classification. In this paper, after image processing, a Mask R-CNN model is built for lesion segmentation, where transfer learning is utilized by using the pre-trained weights from Microsoft COCO dataset. The weights of the trained Mask R-CNN model are saved and transferred to the next task – skin lesion classification, to train a Mask R-CNN model for classification. Our experiments are conducted on the benchmark datasets from the International Skin Imaging Collaboration 2018 (ISIC 2018) and evaluated by the same metrics used in ISIC 2018. The lesion boundary segmentation and lesion classification have achieved an accuracy of 96% and a balanced multiclass accuracy of 80%, respectively.**

Keywords—skin cancer, dermoscopic images, lesion boundary segmentation, lesion classification, Mask R-CNN, transfer learning, deep learning

I. Introduction

Skin cancer is the most common cancer in the United States and worldwide, which has a high mortality rate. According to the Skin Cancer Foundation, one in five Americans will develop skin cancer by the age of 70, and more than 2 people die of skin cancer in the U.S. every hour [1]. Fortunately, skin cancer may be treated successfully if it can be detected at an early stage. For example, when detected early, the 5-year survival rate for melanoma – the deadliest type of skin cancer is 99 percent [1]. Therefore, early and accurate diagnosis of skin cancer is particularly important.

Unaided visual inspection by dermatologists only has a diagnosis accuracy of about 60% [2]. After the development of dermoscopy and dermoscopic algorithms, such as "chaos and clues" [3], "ABCD rule" [4], and "7-point checklist" [5], the diagnosis accuracy of skin cancer has been much higher than the visual inspection [2, 6]. However, the diagnosis accuracy is highly depending on the experiences of clinicians. Hence, it is desirable to develop approaches for automated segmentation and classification of skin cancer.

In the past few years, deep learning has achieved remarkable success in computer vision. Several deep learning approaches have been proposed for skin lesion segmentation on dermoscopic images, using well-known convolutional neural network (CNN) architectures or novel architectures that are trained from scratch. For example, an improved neural network has been proposed for melanoma segmentation, which includes multiple convolution layers, dropout layer, softmax layer, multiple filters, and activation functions [7]. In addition, a hybrid method has been proposed to use convolutional and recurrent neural networks for melanoma segmentation [8]. An ensemble method based on Mask R-CNN and DeeplabV3+ image segmentation model has been developed for skin lesion segmentation [9]. Ünver et al. [10] combined GrabCut algorithm and a CNN model – YOLO for skin lesion segmentation. Additionally, an approach has been presented for automatic skin lesion segmentation, which is to train a 19-layer deep convolutional neural network using a novel loss function based on Jaccard distance [11].

On the other hand, several methods have been developed for skin lesion classification on dermoscopic images, by training a deep learning model (e.g. CNN [12]) from scratch. For example, Harangi [13] has trained four different deep neural network architectures and then fused the outputs of the classification layers to aggregate CNNs into one framework, for classification of melanoma, nevus, and seborrheic keratosis. Yu et al. [14] have proposed very deep residual networks for automated melanoma recognition. An approach for skin lesion classification using deep CNN and data augmentation has been proposed by Pham et al. [15]. In addition, a deep CNN has been trained to classify skin lesions into benign or malignant based on a novel regularizer [16].

Training a deep learning model from scratch requires a large number of labeled data to achieve a good performance. However, in the field of medical image analysis, the number of images with ground truth labels is usually limited. To overcome the limited dataset issue, transfer learning has been recently developed to transfer the weights learned from a trained model for a specific application to another application. There are lots of researches on applying transfer learning to skin lesion classification. For example, Mahbod et al. [17] have employed three different pre-trained CNNs, followed by a multi-scale multi-CNN fusion approach for skin lesion classification on cropped dermoscopic images of various sizes. An approach has been proposed to apply transfer learning to the Alex-net and conduct data augmentation, for skin lesion classification into melanoma, nevus, and seborrheic keratosis [18]. Lopez et al. [19] have made use of a VGGNet pre-trained on ImageNet by training only the higher-level portion of the convolutional layers for skin lesion classification. On the other hand, some researches have employed transfer learning for skin lesion segmentation. For instance, transfer learning has been employed to train a CNN model for skin lesion segmentation, by utilizing pre-trained weights trained on the ImageNet dataset [20].

When training a model for skin lesion classification, the existing researches often do not take lesion segmentation into consideration. That is to say, the model for skin lesion segmentation and the model for skin lesion classification have no connections. Nevertheless, knowledge learned from lesion segmentation, such as borders of lesions, should be transferred to the skin lesion classification task to help train the model.

In our work, due to the limited number of images with labeled masks in the public dataset International Skin Imaging Collaboration 2018 (ISIC 2018) [21, 22], we train a Mask R-CNN model for skin lesion segmentation, by utilizing the pre-trained weights obtained via training on Microsoft COCO dataset [23]. Weights of the trained Mask R-CNN model are saved. We then again train a Mask R-CNN model for lesion classification, by using the pre-trained weights obtained from lesion segmentation. Hence, the knowledge learned from lesion segmentation is transferred to the task of skin lesion classification. ISIC 2018 datasets include a total of 2,594 dermoscopic images with corresponding 2,594 binary mask images for skin lesion segmentation, and a total of 10,015 dermoscopic images of 7 different types of skin lesions for skin lesion classification. Image preprocessing is conducted before training the models, including cropping and resizing, color correction, contrast adjustment, hair removal, and denoising. The experiments have achieved an accuracy of 96% for lesion segmentation, and a balanced multiclass accuracy of 80% for lesion classification.

The paper is organized as follows. In section II, we describe the dataset and the approach used in this work, including training a Mask R-CNN with transfer learning for lesion segmentation, and training a Mask R-CNN for lesion classification with transfer learning using the weights learned from lesion segmentation. Section III presents the experimental results of lesion segmentation and lesion classification. In Section IV, we draw conclusions and discuss the future work.

II. Materials and Methods

A. Dataset

The benchmark datasets provided by ISIC 2018 [21, 22] are used in our work, including one dataset for lesion boundary segmentation and another dataset for lesion diagnosis. Images are taken from various parts of the human body under several differently controlled environmental conditions from different institutions. Since the test data provided by ISIC 2018 do not have ground truth available for both lesion segmentation and classification, we use the provided training datasets and their ground truth as the entire datasets in our experiments.

For the task of lesion segmentation, the benchmark dataset includes a total of 2,594 dermoscopic lesion images, and the corresponding 2,594 binary mask images. The lesion images are color images in JPG format, which contain exactly one primary lesion. The sizes of lesion images vary. The mask images are grayscale PNGs, which have the same dimensions as their corresponding lesion images. The mask images indicate the locations, shapes and borders of the primary skin lesions. Figure 1 shows examples of lesion images and the corresponding mask images in ISIC 2018.

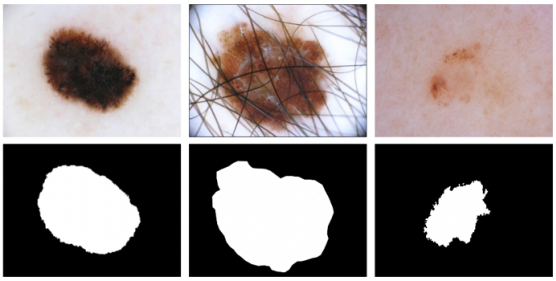

Figure 1 Examples of lesion images (top row) and corresponding mask images (bottom row) in ISIC 2018.

The benchmark dataset for the lesion classification task has a total of 10,015 color dermoscopic images in JPG format, and a CSV file of ground truth indicating one of 7 different diseases for each lesion image. All images are of the same size – 600×450 pixels. The 7 different skin lesions are listed in Table 1, which shows the name, abbreviation, and the number of images of each type of skin lesion. Examples of lesion images for each type of lesions are shown in Figure 2.

Table 1 The number of images of each type of skin lesion in ISIC 2018 for lesion classification.

Type of skin lesion (abbreviation is in the parentheses)	Number
Melanoma (MEL)	1113
Melanocytic Nevus (NV)	6705
Basal Cell Carcinoma (BCC)	514
Actinic Keratosis / Bowen's Disease (AKIEC)	327
Benign Keratosis (BKL)	1099
Dermatofibroma (DF)	115
Vascular Lesion (VASC)	142

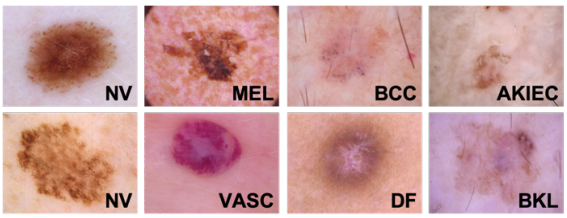

Figure 2 Examples of dermoscopic images of each type of skin lesion in ISIC 2018. Two images of NV are shown here to make the grid filled up.

B. Image Preprocessing

Preprocessing steps include cropping, resizing, correcting color, increasing contrast, removing hair, and reducing noise. Figure 3 shows the pipeline of image preprocessing and the sample output of each step. Details of each preprocessing step are described as follows.

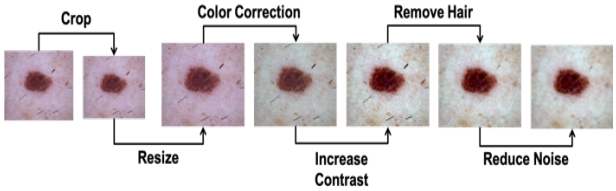

Figure 3 The image preprocessing pipeline in our work

a) Cropping and Resizing

All images are cropped by 2.5% for all four sides of the image. Then all images are resized to a height of 800 pixel. The width is resized using the same ratio as the one used in height resizing.

b) Color Correction

In ISIC 2018 datasets, the dermoscopic images are taken in several different acquisition conditions. To minimize the influence of different acquisition setup on color features extracted from the images, the color compensation technique – Shades of Gray method [24] is applied. The color compensation technique involves two steps: (1) estimation of the color of the light source in RGB coordinates, and (2) transformation of the images using the estimated illuminant [24]. The parameter p (the degree of the norm) is set to 6 and 3, for the lesion segmentation and lesion classification tasks, respectively.

c) Contrast Adjustment

The contrast of all images is adjusted by a contrast adjustment factor which is calculated using the following formula [25].

$$F = \frac{259 \times (L + 255)}{255 \times (259 - L)} \qquad (1)$$

The value L in the formula (1) denotes the desired level of contrast and F is the contrast adjustment factor. L is set to 50 in the preprocessing step to increase contrast without losing too many details of the images. After calculating the contrast adjustment factor F, each pixel's color value is recalculated by the following formula and truncated so that the new color value is in the range of 0 to 255 [25]. In formula (2), C denotes a pixel's original color value and C' denotes the new color value. The formula (2) is applied to the red, green, blue component of a pixel's color separately.

$$C' = F \times (C - 128) + 128 \qquad (2)$$

d) Hair Removal

The dermoscopic images in ISIC 2018 datasets often have hairs. The presence of hair in images may add noises to the images and interfere with the subsequent model training process, hence we employ the hair removal algorithm proposed by [26] to remove hairs. The algorithm consists of two steps: (1) hair detection with the use of adaptive canny edge detector and refinement by morphological operators and (2) hair repair by multi-resolution coherence transport inpainting technique [26].

e) Denoising

To reduce random noise of dermoscopic images, OpenCV python package cv2's 'fastNlMeans-DenoisingColored' function is applied with filtering strength for luminance component h = 5, *templateWindowSize* = 7, and *searchWindowSize* = 21.

C. Lesion Segmentation and Lesion Classification

For lesion segmentation, we train a Mask R-CNN using the weights pre-trained on COCO dataset. The weights are loaded from Model ZOO. To utilize the pre-trained model on COCO dataset, we transform the lesion images into COCO format. The combination of ResNet101 and feature pyramid network (FPN) is used as the backbone of the Mask R-CNN. When training the model for lesion segmentation, we proceed just like training for COCO dataset except that the last layer output units take two classes – lesion or background, instead of 81 classes (80 classes for object categories in COCO dataset and 1 class for background) in the pre-trained model. The weights of the newly trained model are saved. We then preprocess the images in the lesion classification dataset, transform them into COCO format, and train a Mask R-CNN using the pre-trained weights obtained from image segmentation. Training the model for lesion classification is similar to the training for image segmentation, but the last layer output units take 8 classes, including 7 classes for 7 types of skin lesions in the benchmark dataset and one built-in class for indicating the image background. In the Mask R-CNN, region proposal network (RPN) generates region of interests (ROIs), which are aligned with box heads to predict class (lesion or background for lesion segmentation; lesion type or background for lesion classification) and boundary box detections (a set of predicted boundaries). The boundary box detection results align with mask heads to generate the final mask.

GPU provided by free tier of Google Colab is used to train the models. We set batch size = 2 and base learning rate = 0.0025, applying *'linear learning rate scaling rule'* based on Google Colab's environment specifications. The ROI head batch size is set to 128 per image. Also, the maximum iteration size is 2,500 and 14,000, for the lesion segmentation task and the lesion classification task, respectively. Due to the limited computing power on the free tier of Google Colab, the epoch size is set to 2.

Ten-fold cross validation is used to evaluate both lesion segmentation and lesion classification. In the task of lesion segmentation, the 2,594 lesion images and their corresponding 2,594 mask images are divided into training and testing datasets with a 9 to 1 ratio. For lesion classification, we conduct up-sampling and down-sampling to balance the number of images in each class. This is to resolve the class imbalance issue reflected in Table 1 - Melanocytic Nevus (NV) images take over 66% of the entire dataset, and Dermatofibroma (DF) takes less than 1.2% of the dataset. We up-sampled data to 2,000 for each class except NV by copying randomly selected images. The number of NV is down-sampled to 2,000 by randomly select 2,000 images from the 6,705 images of NVs. Through up-sampling and down-sampling, a total of 14,000 images is generated. We randomly select 1,400 images to serve as the testing dataset, and the other 12,600 images are used as the training dataset.

III. Results

A. Evaluation Metrics

We use the same evaluation metrics as the metrics that are used in ISIC Challenge 2018. For lesion segmentation, evaluation is done in pixel-level. Accuracy, sensitivity, specificity, Dice Coefficient, Jaccard Index, and Threshold Jaccard Index are used to evaluate the model, which formulas are listed below in (3) – (7). Threshold Jaccard Index uses the same formula as the formula (7), with one exception – if the calculated Jaccard Index falls below a threshold T, the Jaccard Index is set to zero; otherwise, use the calculated value as the Jaccard Index. In our experiment,

the threshold T was set to 0.65, the same as the threshold used in ISIC 2018 Challenge.

The evaluation of lesion classification is conducted at the whole image level. The metrics for lesion classification include accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), Dice Coefficient, and Balanced Accuracy. The formulas of PPV and NPV are listed below in (8) and (9). Balanced Accuracy is the mean recall across classes after mutually exclusive classification decision.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{Dice Coefficient} = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (6)$$

$$\text{Jaccard Index} = \frac{TP}{TP + FN + FP} \quad (7)$$

$$\text{PPV} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{NPV} = \frac{TN}{TN+FN} \quad (9)$$

In the formulas, *TP, FP, TN, FN* denote the number of true positives, false positives, true negatives, and false negatives, respectively.

B. Lesion Segmentation Result

Our approach on lesion segmentation has achieved a performance of 0.79, 0.82, 0.89, 0.91, 0.97, and 0.96, for Threshold Jaccard Index, Jaccard Index, Dice Coefficient, sensitivity, specificity, and accuracy, respectively. Figure 4 demonstrates the segmentation result of a sample dermoscopic image.

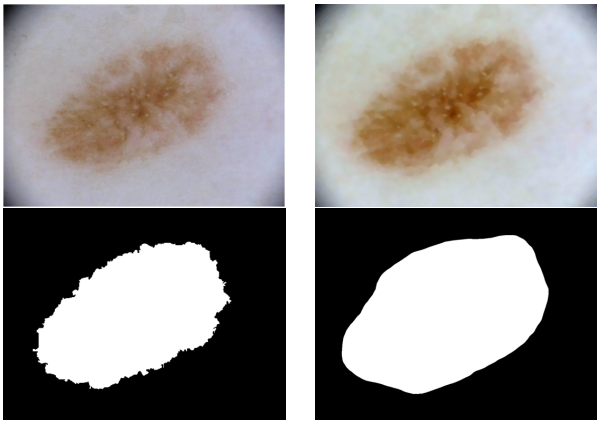

Figure 4 Segmentation result of one sample image. The top left is the original dermoscopic image; the top right is the preprocessed image; the bottom left is the benchmark mask of the image; the bottom right is the lesion segmentation result.

C. Lesion Classification Result

The evaluation results of lesion classification are shown in Table 2, which shows the detailed performance of diagnosing each type of skin lesions (accuracy, sensitivity, specificity, positive predicted value, negative predicted value, and Dice Coefficient are abbreviated as AC, SE, SP, PPV, NPV, and DC, respectively)

Table 2 The performance of our approach for lesion classification. The 1st row lists the abbreviations of each skin lesion, the 1st column is the metrics, and the 2nd column is the average performance (e.g. average accuracy, average sensitivity, etc.). The average sensitivity is the value of Balanced Accuracy.

	Avg	MEL	NV	BCC	AKIEC	BKL	DF	VASC
AC	0.94	0.91	0.93	0.95	0.95	0.87	1.00	1.00
SE	0.80	0.48	0.70	0.67	0.87	0.88	0.99	1.00
SP	0.97	0.97	0.97	1.00	0.96	0.87	1.00	1.00
PPV	0.83	0.72	0.82	0.98	0.81	0.52	0.98	1.00
NPV	0.97	0.92	0.95	0.95	0.98	0.98	1.00	1.00
DC	0.80	0.58	0.76	0.80	0.84	0.65	0.98	1.00

IV. Discussions and Conclusions

In this paper, we present an approach that utilizes transfer learning and deep learning to train a Mask R-CNN for lesion segmentation on preprocessed dermoscopic images and then uses transfer learning again to transfer the weights learned from lesion segmentation to train a Mask R-CNN for lesion classification. Our approach has achieved promising results evaluated by the metrics used in ISIC 2018. To conduct the experiments, we need benchmark datasets for both the lesion segmentation and lesion classification. Hence, we use datasets of ISIC 2018 Challenge in our work because it is the latest ISIC dataset that has both the lesion segmentation and classification benchmark datasets (ISIC 2019 Challenge only provides benchmark dataset for lesion classification but not for lesion segmentation; ISIC 2020 Challenge only provides benchmark dataset for detecting whether the lesion in the image is malignant or benign but not for lesion segmentation and classification of the types of lesions).

In the future, data augmentation and external dermoscopic images will be used to handle a class imbalance issue of different types of skin lesions in the datasets for data classification, instead of using pseudo up-sampling and down-sampling. In addition, the model is expected to be improved with better computing power by training more iterations with optimized hyperparameters of the model.

References

[1]. "Skin Cancer Facts & Statistics." *The Skin Cancer Foundation*, https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/, accessed in June, 2021.

[2]. Kittler, H., Pehamberger, H., Wolff, K., Binder, M.: "Diagnostic accuracy of dermoscopy". *The Lancet Oncology*. vol. 3, no. 3, pp. 159.165. 2002.

[3]. Rosendahl, C., Cameron, A., McColl, I., Wilkinson, D.: "Dermatoscopy in routine practice: 'chaos and clues'". *Aust Fam Physician.* 2012;41(7):482-487.

[4]. Stolz, W., Riemann, A., Cognetta, A.B., et al.: "ABCD rule of dermoscopy: a new practical method for early recognition of malignant melanoma". *Eur J Dermatol*. 1994;4(7):521-527.

[5]. Argenziano G, Fabbrocini G, Carli P, Giorgi V, Sammarco E, Delfino M: "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis". *Arch Derm 1998*, 134:1563–1570.

[6]. Abder-Rahman, A.A., Deserno, T.M.,: "A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data". In *Proc. SPIE 8318, Medical Imaging 2012:*

Image Perception, Observer Performance, and Technology Assessment, 2012.
[7]. Xiaoqing Zhang, “Melanoma segmentation based on deep learning”, *Computer Assisted Surgery*, 22:sup1, 267-277, DOI: 10.1080/24699322.2017.1389405
[8]. M. Attia, M. Hossny, S. Nahavandi and A. Yazdabadi, "Skin melanoma segmentation using recurrent and convolutional neural networks," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Melbourne, VIC, 2017, pp. 292-296, doi: 10.1109/ISBI.2017.7950522.
[9]. M. Goyal, A. Oakley, P. Bansal, D. Dancey and M. H. Yap, "Skin Lesion Segmentation in Dermoscopic Images With Ensemble Deep Learning Methods," in *IEEE Access*, vol. 8, pp. 4171-4181, 2020, doi: 10.1109/ACCESS.2019.2960504.
[10]. Ünver, Halil Murat, and Enes Ayan. "Skin lesion segmentation in dermoscopic images with combination of YOLO and grabcut algorithm." *Diagnostics* 9.3 (2019): 72.
[11]. Y. Yuan, M. Chao and Y. Lo, "Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks With Jaccard Distance," in *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1876-1886, Sept. 2017, doi: 10.1109/TMI.2017.2695227.
[12]. Brinker, Titus Josef, et al. "Skin cancer classification using convolutional neural networks: systematic review." *Journal of medical Internet research* 20.10 (2018): e11936.
[13]. Harangi, Balazs. "Skin lesion classification with ensembles of deep convolutional neural networks." *Journal of biomedical informatics* 86 (2018): 25-32.
[14]. Yu, Lequan, et al. "Automated melanoma recognition in dermoscopy images via very deep residual networks." *IEEE transactions on medical imaging* 36.4 (2016): 994-1004.
[15]. Pham, Tri-Cong, Chi-Mai Luong, Muriel Visani, and Van-Dung Hoang. "Deep CNN and data augmentation for skin lesion classification." In *Asian Conference on Intelligent Information and Database Systems*, pp. 573-582. Springer, Cham, 2018.
[16]. M. A. Albahar, "Skin Lesion Classification Using Convolutional Neural Network With Novel Regularizer," in *IEEE Access*, vol. 7, pp. 38306-38313, 2019, doi: 10.1109/ACCESS.2019.2906241.
[17]. Mahbod, Amirreza, et al. "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification." *Computer Methods and Programs in Biomedicine* (2020): 105475.
[18]. Hosny, Khalid M., Mohamed A. Kassem, and Mohamed M. Foaud. "Classification of skin lesions using transfer learning and augmentation with Alex-net." *PloS one* 14.5 (2019): e0217293.
[19]. Lopez, Adria Romero, Xavier Giro-i-Nieto, Jack Burdick, and Oge Marques. "Skin lesion classification from dermoscopic images using deep learning techniques." *In 2017 13th IASTED international conference on biomedical engineering (BioMed),* pp. 49-54. IEEE, 2017.
[20]. Zafar, Kashan, et al. "Skin Lesion Segmentation from Dermoscopic Images Using Convolutional Neural Network." *Sensors* 20.6 (2020): 1601.
[21]. Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern: “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)”, 2018; https://arxiv.org/abs/1902.03368
[22]. Tschandl, P., Rosendahl, C. & Kittler, H. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. *Sci. Data* 5, 180161 doi:10.1038/sdata.2018.161 (2018).
[23]. Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In *European conference on computer vision*, pp. 740-755. Springer, Cham, 2014.
[24]. Barata, Catarina, Jorge S. Marques, and M. Emre Celebi. "Improving dermoscopy image analysis using color constancy." In *2014 IEEE International Conference on Image Processing (ICIP),* pp. 3527-3531. IEEE, 2014.
[25]. Fatima, et al. “Image Processing Algorithms Part 5: Contrast Adjustment.” *Dreamland Fantasy Studios*, www.dfstudios.co.uk/articles/programming/image-programming-algorithms/image-processing-algorithms-part-5-contrast-adjustment/, accessed in January, 2021.
[26]. Toossi, Mohammad Taghi Bahreyni, et al. "An effective hair removal algorithm for dermoscopy images." *Skin Research and Technology* 19.3 (2013): 230-235.