Deep Skin Lesion Analysis via Dense U-Net with Adaptive Color Augmentation

Anindya Shaha, Prem Prasad, Abdullah Thabit

***Abstract*—Fully automatic detection, segmentation and classification of skin lesions in dermatoscopic images can facilitate early diagnosis and repression of malignant melanoma and non-melanoma skin cancer. In this paper, we present a densely connected convolutional neural network as a powerful solution to tackle this challenge, while drawing a detailed comparative analysis of the fundamental complexities faced during inference. Additionally, we propose a revised, state-of-the-art adaptive color augmentation technique to amplify data expression and model performance. The overall system achieves a Dice Ratio of 0.891 with 0.943 sensitivity and 0.932 specificity, on the ISIC 2018 Testing Set for segmentation. Preliminary cross-validation results indicate 0.72 mean precision and 0.66 mean recall across the 7 different classes included in the ISIC 2018 Training Set for classification.**

***Index Terms*—dermatoscopy, melanoma, convolutional neural network, deep learning, color augmentation, segmentation, classification**

I. Introduction

HUMAN skin, as the largest organ of the body's integumentary system, is prone to a wide spectrum of cutaneous diseases and infections that can manifest as surface abnormalities or "lesions". In particular skin cancer, primarily non-melanoma skin cancer (NMSC) and the highly aggressive malignant melanoma (MM), represents the most common malignancy in Caucasians [1] with over 1.3 million new cases and $125,000$ deaths worldwide in 2018 [1, 2]. Early detection and diagnosis of skin lesions is critical to ensuring high survival rates [3, 9]. It can be achieved effectively, automatically and in real-time, even in the absence of medical expertise, by leveraging complex machine learning techniques and computer vision frameworks [4]. With the emergence of large, multi-source dermatoscopic image datasets [5] providing ample annotated training data, deep neural networks are now at the forefront of this technology as universal approximators.

In this research, we propose a deep convolutional neural network (CNN) with dedicated branches to segment and classify seven distinct classes of the most commonly occurring pigmented skin lesions. Additionally, we incorporate a novel adaptive color augmentation technique, with improved functionality from its equivalent counterparts [6, 7], to extend our training data representation. The augmentation exploits and accounts for the highly variable nature of dermatoscopic screening samples, where background illumination, hospital acquisition conditions and external obstructions can significantly modify the underlying color profile of a skin lesion captured in an image. Performance analysis is modelled after the *ISIC 2018 Challenge: Skin Lesion Analysis for Melanoma Detection* [8] using the HAM10000 [5] dataset.

A. Shaha, P. Prasad, A. Thabit are Erasmus+ Joint Master in Medical Imaging and Applications (MaIA) candidates, under the consortium of Universitat de Girona (Spain), Università degli Studi di Cassino (Italy) and Université de Bourgogne (France). Contact: anindo@ieee.org, premprasath95@gmail.com, abdullah.thabitt@gmail.com.

II. Artificial Data Augmentation

HAM10000 depicts 7 types of skin lesions (melanoma [MEL], melanocytic nevus [NV], basal cell carcinoma [BCC], actinic keratosis [AKIEC], benign keratosis [BKL], dermatofibroma [DF], vascular lesion [VASC]) in $10,015$ images. However, a complete visual representation of these classes demands a much larger, unfeasible number of images, i.e. theoretically every possible instance in nature. The most practical means of compensation is to anticipate and adapt near-realistic variations of the images beyond the pre-existing, limited dataset using data augmentation. This step is proven to have a significant impact on inference by reducing overfitting and improving generalization [9].

A. Color Augmentation

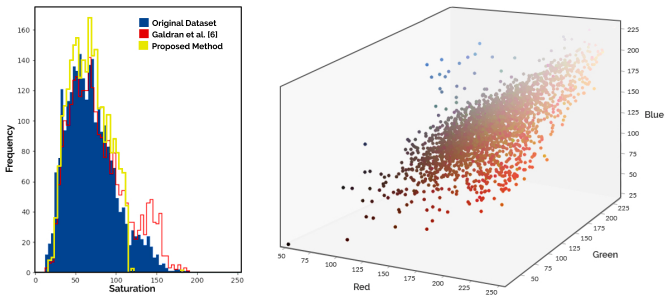


Figure 1.Histogram of mean saturation values across the original training set and the augmented datasets, derived from the coordinates of each image in the HSV color space (left); Distribution of illuminant profiles extracted from the training set by applying the *Gray World* color constancy algorithm, where each marker is displaying its encoded color in the RGB color space (right).

Color is an important feature for diagnosing malignant melanoma since certain color markers are associated with different classes of melanoma [9]. A popular approach to achieve color constancy is the *Gray World* algorithm that assumes the average color in a given scene is achromatic, i.e. gray, and any deviations is caused due to effects of light sources [10]. Based on this, the illuminant profile of an image can be estimated as the independent average intensities of its RGB color channels. In turn, the scaling factor for each

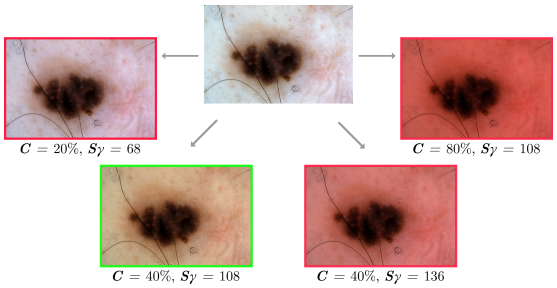


Figure 2. Possible color augmentations for an image and their respective values for C and S_γ, with the negative candidates marked in red and the positive candidate (fulfilling both pre-requisite conditions) marked in green.

channel $(\beta_R, \beta_G, \beta_B)$ is its average intensity divided by that of all 3 channels combined. Together, these scaling factors constitute as the illuminant scales (β') required to transform any image to a certain illuminant profile (β), given as follows:

$$\beta'\{\beta_R, \beta_G, \beta_B\} \equiv \beta'\{\frac{\frac{1}{n}\sum I_R}{\sum I_{RGB}}, \frac{\frac{1}{n}\sum I_G}{\frac{1}{3n}\sum I_{RGB}}, \frac{\frac{1}{n}\sum I_B}{\frac{1}{3n}\sum I_{RGB}}\}$$

Dividing the original image (I) by its illuminant scales (β'), results in a white-balanced image. Similarly, if we take the illuminant scales of a different image (γ') and multiply it by the original white-balanced image, we obtain a color augmented image (I_{aug}) with the illuminant γ.

$$I_{aug} = (I/\beta') * \gamma'$$

This method was originally proposed by Galdran *et al.* [6], but faces important limitations. The authors perform color augmentations at train-time, casting each sample by an illuminant profile (with uniform probability distribution) that is randomly selected from the raw empirical distribution of all illuminants present in the dataset. This leads to high variance while training the network, as a training set of 2072 images can yield 2072^2 possible variants, where each image is learned to different degrees, at random. In certain cases, augmented images can also exhibit an oversaturated, artificial hue [6], by adopting an illuminant profile vastly different from its original. These images can prove detrimental to the learning process. In this paper, we propose an updated approach to account for these limitations.

Fig. 1(b) illustrates how the overall illuminant distribution of the dataset is centered around a reddish hue, with a few outliers at both extremes. Saturation values follow accordingly, with a major Gaussian distribution centered around 60 and a minor branch near 140. The secondary distribution represents highly saturated images in the original dataset and its proportion is boosted dramatically in a color augmented dataset generated by the original algorithm [6], as seen in Fig. 1(a). Taking these factors into account, we introduce two strict conditions prior to augmentation:

$$d(\beta, \gamma) \approx C * d(\beta, \alpha) \quad \wedge \quad S_\gamma \in [a, b]$$

Here $d(\beta, \alpha)$ represents the Euclidean distance between the illumination profile (β) of an image (I) and its furthest counterpart (α) in the RGB color space. $d(\beta, \gamma)$ represents the equivalent between β and the candidate illuminant to be used for augmentation (γ). C is used as a thresholding factor, set as 0.4. In other words, if γ is sufficiently distinct (40%) from β, yielding new trainable data), but not as radically different as α such as to impair learning, we consider it as a positive candidate for augmentation. For the second condition, S_γ represents the mean saturation value of I_{aug} (post-augmentation with γ). If S_γ pertains to the major distribution range $[a, b]$ of saturation in the original dataset, then we confirm the augmentation and use it for training. Otherwise, we iterate to the next candidate illuminant satisfying the first condition and verify the same. Fig. 2 illustrates this selection process. The values of a and b are determined to be 15 and 115, respectively, for HAM10000. As a result, the algorithm effectively eliminates the generation of oversaturated, highly artificial images during augmentation (refer to Fig. 1(a)).

B. Spatial Augmentation

Morphological transformations, such as rotation ($-180°$ to $180°$), horizontal/vertical flip and translation (10% along x, y-axis), are used to create spatial augmentations at train-time and account for learning orientations beyond the dataset.

III. Network Architecture

A. Deep Neural Network for Lesion Segmentation

TABLE I: Hyperparameter Tuning and Optimization

Hyperparameters	Tuning Range	Optimal Values
Learning Rate (α)	$10^{-15} - 0.1$	$\mathbf{10^{-7} - 10^{-4}}$
α Decay Factor	$0.05 - 0.5$	**0.5**
α Decay Patience	$1 - 10$	**2**
Mini-Batch Size	$6 - 32$	**16**
Optimizer Type	SGD, Adam, RMSprop	**Adam**
Optimizer Parameter (β_1)	$0 - 1$	**0.9**
Optimizer Parameter (β_2)	$0 - 1$	**0.999**
Encoder Initialization	ImageNet[1], He, Xavier	**ImageNet[1]**
Decoder Initialization	He, Xavier, Normal	**Xavier**
Encoder Activation	ReLU, LReLU, $tanh$	**ReLU**
Decoder Activation	ReLU, LReLU, σ	**ReLU[2]**
Trainable Encoder Layers	$1 - 201$	**201**
Trainable Decoder Layers	$1 - 4$	**4**
Training Period	$5 - 200$	**7**

[1] Pre-trained weights for ILSVRC ImageNet database [11].
[2] σ activation is used in output layer.

The base architecture is a standard U-Net, as proposed by O. Ronneberger *et al.* [12] for biomedical image segmentation. It comprises of a backbone encoder (series of downsampling convolutional layers used for feature extraction) followed by a decoder (corresponding number of upsampling transposed convolutional layers) to deliver pixel-level classification in an output segmentation map of the original input size. All images are normalized and pre-processed to 224×224 pixels for ease of computation and uniformity through the hyperparameter tuning and optimization phase, and these dimensions

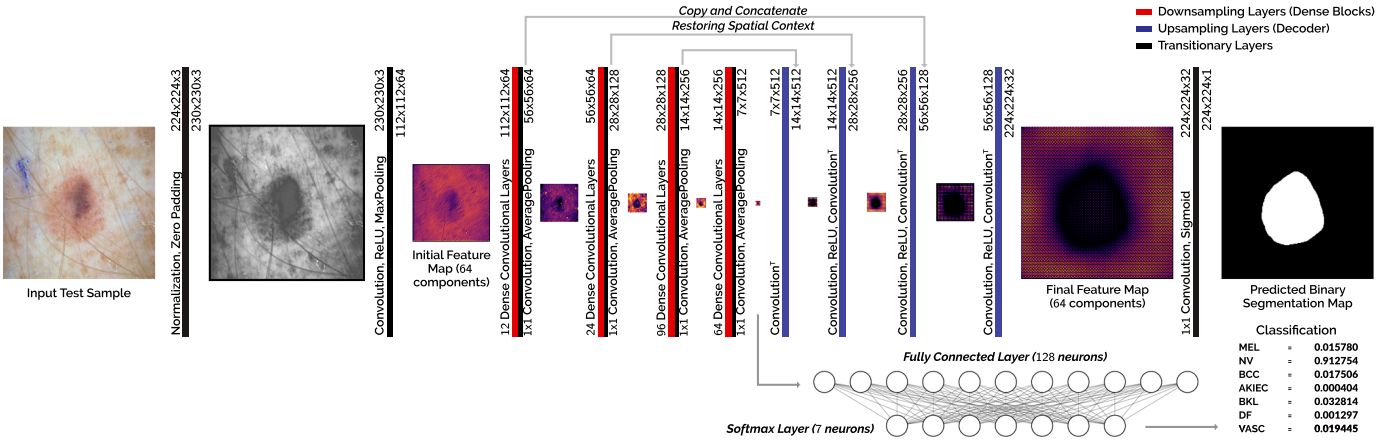


Figure 3. Complete architecture for proposed U-Net, with $DenseNet201$ backbone and dedicated branches for segmentation and classification stemming from the output of the final downsampling layer. Every dense block is made of several alternating 1×1 and 3×3 convolutional layer pairs.

serve as the input size for the network. Rectified Linear Unit (ReLU) is used to activate all convolutional layers with the exception of the output layer, where the sigmoid (σ) function has been used instead. Binary cross-entropy loss and Jaccard Index are used as the primary evaluation metrics with gradient descent and Adam optimizer in backpropagation. To determine the most appropriate backbone for this network, a comparative analysis is drawn across several notable architectures ($ResNet50$, $VGG16$, $VGG19$, $DenseNet169$, $DenseNet201$, $Inception(v3)$), each independently combined with the U-Net in separate turns. They are tested for 2072 training samples and 518 validation samples, over at training period of 25 epochs, mini-batch size of 32 and 5-fold Monte Carlo cross validation. Due to its superior performance, $DenseNet201$ is selected as the encoder in the final model. $DenseNet$ layers take the concatenated feature maps of all preceding layers as their input and similarly pass on their output feature map to all subsequent layers [13]. As a result, the model requires less channels, is computationally/memory-wise efficient, has strong gradient flow during backpropagation and considers diverse features of different complexities at every stage of computation. The complete U-Net is then tuned across its hyperparameters and trained to maximize Jaccard Index over 5-fold Monte Carlo cross validation. The optimal operating values have been noted in Table I.

B. Deep Neural Network for Lesion Classification

An auxiliary branch connected at the end of the downsampling $DenseNet201$ feature extractor enables the U-Net to perform classification, as shown in Fig. 3. Although fully functional, this branch is still under development for tuning and optimal performance. Currently, it comprises of a fully connected layer with 128 neurons, followed by a softmax layer with 7 neurons for the 7 different categories to be classified. It operates independently from the segmentation branch at train-time, however reuses the features learned from its backbone. Training performance in this stage is evaluated using categorical cross-entropy loss, precision and recall.

IV. Experimental Results

The model is implemented using the high-level neural network API, Keras with TensorFlow as the backend engine. A single Nvidia Tesla K80 GPU (12 GB) is used for hardware acceleration via Google Colaboratory cloud service. After training on the complete annotated *ISIC 2018 Task 1: Training Set*, the final segmentation results from inference on *ISIC 2018 Task 1: Testing Set* have been recorded in Table II. Post-proccessing includes simply smoothing and extracting the largest connected component from the predicted binary mask. Similarly, preliminary 5-fold cross validation results for classification on *ISIC 2018 Task 3: Training Set* have been recorded in Table III.

TABLE II: Segmentation Performance for Testing Set (JA: Jaccard Index; DI: Dice Ratio; AC: Accuracy; SE: Sensitivity; SP: Specificity)

Method	JA	DI	AC	SE	SP
Yuan et al. [14][1]	0.765	0.849	0.934	0.825	0.975
Sarker et al. [15][1]	0.782	0.878	0.936	0.816	**0.983**
Venkatesh et al. [16][1]	0.764	0.856	0.936	0.830	0.976
Galdran et al. [6][1]	0.767	0.846	**0.948**	0.865	0.980
Proposed Dense U-Net[2]	**0.819**	**0.891**	0.937	**0.943**	0.932

[1] Training:Test Ratio = 2000:600 (ISIC 2017)
[2] Training:Test Ratio = 2594:1000 (ISIC 2018)

TABLE III: Classification Performance for Training Set (5-Fold Cross Validation; Training:Test Ratio = 8012:2003)

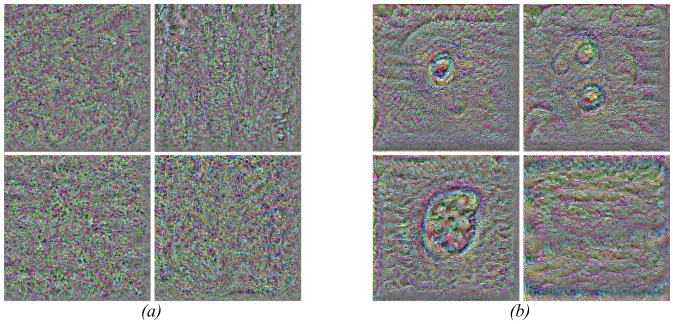
Metric	MEL	NV	BCC	AKIEK	BKL	DF	VASC	Mean
Precision	0.58	0.92	0.77	0.71	0.59	0.57	0.89	**0.72**
Recall	0.56	0.92	0.94	0.38	0.73	0.55	0.83	**0.66**

V. Discussion

A. Understanding Segmentation Features

Initializing the $DenseNet201$ encoder with pre-trained weights from ImageNet database provides a large boost in performance, as the network benefits from a strong understanding of base features built upon 10 million images, as

well as faster convergence times [17]. However, this step can be further extended. After the U-Net has been successfully trained to perform segmentation, we are left with a network that not only recognizes base features derived from ImageNet, but now holds resultant weights ($SLweights$) from further training on HAM10000, thereby targeting skin lesion features. This can be verified by passing a monochrome noisy image to the model, monitoring the activation of a specific filter within a specific layer, and performing gradient ascent to modify the input image in order to maximize the filter activation [18]. By doing so, we can identify and visualize the input image or features (refer to Fig. 4) that the network is sensitive to and has learned throughout its hierarchy via training. $SLweights$ can then be used to initialize the encoder during the classification stage, instead of standard ImageNet weights. A similar approach was demonstrated by Wong et al.[19], where the authors successfully illustrated the advantages of utilizing segmentation features for classification, by freezing the output of the segmentation network and feeding it to a series of convolutional blocks. Alternately, both branches can also be trained simultaneously and evaluated using a combined loss function to achieve generalized joint proficiency in both segmentation and classification.



(a) *(b)*

Figure 4. Images generating maximum activation for filters in the final dense block convolutional layer, when trained on: a) ImageNet −generalized texture features based on 10 million images and 1000 classes; b) HAM10000 −specialized features based on 2594 dermatoscopic images and 7 classes.

B. Learning Effect of Augmentations

Color and spatial augmentation is able to provide a combined boost of approximately 2% in segmentation. However, their full potential and effect in classification remains to be studied. One particular challenge involves determining the appropriate range of learning rates for this artificially enlarged, dynamic dataset. Cyclic learning rates [20], as opposed to a global, monotonically decaying learning rate, may be a possible solution to address this problem.

VI. Conclusion

In conclusion, an end-to-end integrated densely connected convolutional neural network is presented for segmenting and classifying skin lesions in dermatoscopic screening images. Furthermore, a data-driven adaptive color augmentation technique is redesigned and improved, yielding significant performance increase for the model. Considering the multiple focus areas that can be investigated in future research and the ever-evolving properties of deep learning architectures, prospects of a revised model with notable improvements are highly favorable. The promising results of this research verify the notion and feasibility of realizing a day when fast, automated timely diagnosis of skin cancer, without the necessity of medical supervision, will be easily accessible and reliable.

References

[1] Bray, F. et al. (2018), "Global Cancer Statistics 2018", *CA: A Cancer Journal for Clinicians*, vol. 68:6, pp. 394–424. DOI:10.3322/caac.21492

[2] Apalla, Z. et al. (2017), "Skin Cancer: Epidemiology, Disease Burden, Pathophysiology, Diagnosis, and Therapeutic Approaches", *Dermatology and Therapy*, vol. 7:S1, pp. 5–19. DOI:10.1007/s13555-016-0165-y

[3] Rigel, D., Carucci, J. (2000), "Malignant Melanoma: Prevention, Early Detection, and Treatment in the 21st Century", *CA: A Cancer Journal for Clinicians*, vol. 50:4, pp. 215–236. DOI:10.3322/canjclin.50.4.215

[4] Frost & Sullivan Research (2001), "U.S. Emerging Melanoma Therapeutics Market". Retrieved from: http://www.frost.com/sublib/display-report.do?id=A090-01-00-00-00

[5] Tschandl, P., Rosendahl, C., Kittler, H. (2018), "The HAM10000 Dataset, A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions", *Nature: Scientific Data*, vol. 5:180161. DOI:10.1038/sdata.2018.161

[6] A. Galdran et al. (2017), "Data-Driven Color Augmentation Techniques for Deep Skin Image Analysis". arXiv:1703.03702[cs.CV]

[7] Z. Lou et al. (2015), "Color Constancy by Deep Learning", *British Machine Vision Conference (BMVC)*, pp. 76.1–76.12. DOI:10.5244/C.29.76

[8] Codella N. et al. (2018), "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)". arxiv:1902.03368[cs.CV]

[9] Stanley, R. J., Stoecker, W. V., Moss, R. H. (2007), "A Relative Color Approach to Color Discrimination for Malignant Melanoma Detection in Dermoscopy Images", *Skin Research and Technology*, vol. 13:1, pp. 62–72. DOI:10.1111/j.1600-0846.2007.00192.x

[10] Finlayson, G.D. et al. (1998), "Comprehensive Colour Normalization", *Proc. European Conf. on Computer Vison (ECCV)*, vol. 1, pp. 475–490.

[11] Russakovsky O. et al. (2015), "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, vol. 115:3, pp. 211–252. DOI:10.1007/s11263-015-0816-y

[12] Ronneberger, O., Fischer, P., Brox, T., (2015), "U-Net: Convolutional Networks for Biomedical Image Segmentation", *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, Springer.

[13] Jegou, S. et al. (2017), "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation", *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. DOI:10.1109/cvprw.2017.156

[14] Yuan, Y. et al. (2017), "Improving Dermoscopic Image Segmentation With Enhanced Convolutional-Deconvolutional Networks", *IEEE Journal of Biomedical and Health Informatics*, vol. 23:2, pp. 519-526. DOI:10.1109/JBHI.2017.2787487

[15] Sarker, M. et al. (2018), "SLSDeep: Skin Lesion Segmentation Based on Dilated Residual and Pyramid Pooling Networks", *Medical Image Computing and Computer Assisted Intervention*, vol. 11071, Springer.

[16] Venkatesh, G.M. et al. (2018), "A Deep Residual Architecture for Skin Lesion Segmentation", *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, vol 11041, Springer.

[17] He, K., Girshick, R., Dollár, P. (2018), "Rethinking ImageNet Pre-Training". arXiv:1811.08883[cs.CV]

[18] Yosinski, J. et al. (2015), "Understanding Neural Networks Through Deep Visualization", *Deep Learning Workshop, 31st International Conference on Machine Learning*. arXiv:1506.06579[cs.CV]

[19] Wong, K. et al. (2018), "Building Medical Image Classifiers with Very Limited Data using Segmentation Networks", *Medical Image Analysis*, vol. 49, pp. 105-116. DOI:10.1016/j.media.2018.07.010

[20] Smith, L., Topin, N. (2017), "Exploring Loss Function Topology with Cyclical Learning Rates", *IEEE International Conference on Learning Representations*. arXiv:1702.04283 [cs.LG]