WHAT EVIDENCE DOES DEEP LEARNING MODEL USE TO CLASSIFY SKIN LESIONS?

Junyan Wu [1,†], Xiaoxiao Li [2,†,⋆], Eric Z. Chen [3,†], Hongda Jiang [4,†], Xu Dong [5] and Ruichen Rong [6]

[1] Cleerly Inc, New York City, NY, USA
[2] Yale University, New Haven, CT, USA
[3] Dana-Farber Cancer Institution, Boston, MA, USA
[4] East China University of Science and Technology, Shanghai, China
[5] Virginia Tech, Blacksburg, VA, USA
[6] UT Southwestern Medical Center, Dallas, TX, USA

ABSTRACT

Melanoma is a type of skin cancer with the most rapidly increasing incidence. Early detection of melanoma using dermoscopy images significantly increases patients' survival rate. However, accurately classifying skin lesions, especially in the early stage, is extremely challenging via dermatologists observation. Hence, discovery of reliable biomarkers for melanoma diagnosis will be meaningful. Recent years, deep learning empowered computer assisted diagnose has been shown its value in medical imaging based decision making. However, lots of research focus on improving the disease detection accuracy but not exploring the evidence of pathology. In this paper, we propose a method to interpret the deep learning classification findings. Firstly, we propose an accurate neural network architecture to classify skin lesion. Secondly, we utilize a prediction difference analysis method that examining the each patch on the image through patchwised corrupting for detecting the biomarkers. Lastly, we validate that our biomarker findings are corresponding to the patterns in the literature. The findings might be significant to guide clinical diagnosis.

Index Terms— Skin lesion, Dermoscopy, Deep learning, Interpretation, Melanoma

1. INTRODUCTION

Skin cancer is a severe public health problem in the United States, with over 5,000,000 newly diagnosed cases every year. Melanoma, as the severest form of skin cancer, is responsible for 75% of deaths associated with skin cancer [1].

Dermoscopy is one of the most widely used skin imaging to distinguish the lesion spots on skin due to its noninvasiveness [2]. Nevertheless, the automatic recognition of melanoma using dermoscopy images is still a challenging task due to the following reasons. First, the low contrast between skin lesions and normal skin regions makes it difficult to accurately segment lesion areas; Second, the melanoma and non-melanoma lesions may have a high degree of visual similarity; Third, the variation of skin conditions such as skin color, natural hairs or veins, among patients produce the different appearance of melanoma, in terms of color and texture, etc. Recent works employed Convolutional Neural Networks (CNNs) have shown its improved discrimination performance in melanoma classification aiming at taking advantage of their discrimination capability to achieve performance gains [3]. Although these studies focused on improving computer assisted diagnostic accuracy, the diagnosis itself is hard even for experienced clinical practitioner based on dermoscopy images. The computer intervention not only forward assist decision making, but it also can benefit clinical research to identify the biomarkers which contribute to diagnosing. Despite promising results, the clinicians typically want to know if the model is trustable and how to interpret the results. Biomarker interpretation from deep learning models for clinical use has been explored in identifying brain disease [4, 5]. However, to the best of our knowledge, the evidence which deep learning model uses for classifying skin lesions has not been explored. Experienced dermatologists diagnose skin diseases based on comprehensive medical criteria which have been verified to be useful, e.g., the ABCD rule [6] and the 7-point checklist [7], etc. We aim to inspect whether the deep learning models and the dermatologists use similar criteria. Motivated by this, in this study, we propose a pipeline to identify the evidence and biomarkers in the skin lesion dermoscopic images, which contributes to the deep learning classifier. To be specific, we first trained an accurate deep learning model to classify each dermoscopic image, with a predicted probability score to each class. Secondly, we analyze the feature importance by corrupting with conditional sampling, then compare the prediction difference.
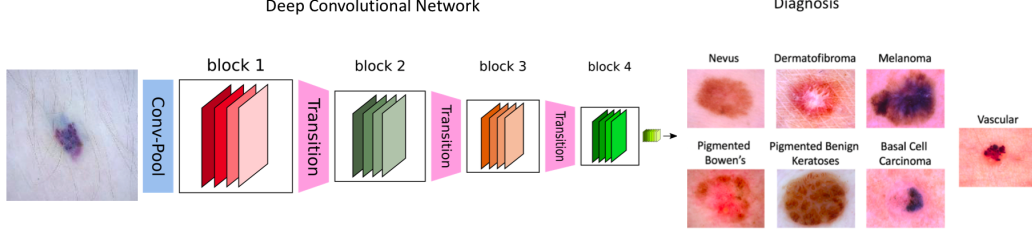
Fig. 1. The flowchart of deep learning model to classify skin lesions [8]

2. METHOD

2.1. Deep Learning Image Classifier

CNNs have been widely used in natural images classification and object recognition, due to its hierarchical feature learning capability and state-of-the-art discrimination performance. For instance, the CNN based methods outperform the traditional techniques significantly in the recent ImageNet challenges [9]. In this work, an image based skin lesion classification task is solved using classifiers based on ResNet and VGG networks. Specifically, the final classification is made by assembling ResNet and VGG networks through lightGBM [10], which aims to improve the classification result obtained by a single classifier.

2.2. Interpreting Deep Learning Features

In order to interpret the feature importance to a classifier, we analysis the classification differences with/without the specific feature, i.e., the difference between $p(c|\boldsymbol{X})$ and $p(c|\boldsymbol{X}_{\backslash i})$, where $\boldsymbol{X}$ represents all the input features and $\boldsymbol{X}_{\backslash i}$ denotes the set of input features except x_i. A large difference indicates the feature contributes significantly to the final decision making, whereas a small difference means the feature is less important to the classification result. In [5], this difference was evaluated by weight of evidence (WE), which is expressed by

$$\text{WE}_i(c|\boldsymbol{X}) = \log_2(\text{odds}(c|\boldsymbol{X})) - \log_2(\text{odds}(c|\boldsymbol{X}_{\backslash i})) \quad (1)$$

where $\text{odds}(c|\boldsymbol{X}) = p(c|\boldsymbol{X})/(1 - p(c|\boldsymbol{X}))$ [1] and

$$p(c|\boldsymbol{X}_{\backslash i}) = \sum_{x_i} p(x_i|\boldsymbol{X}_{\backslash i})p(c|\boldsymbol{X}_{\backslash i}, x_i). \quad (2)$$

In this work, we also use WE to illustrate the importance of a feature x_i with respect to (w.r.t.) a class c. Note that the calculation of WE w.r.t. each feature is unrealistic during implementation due to the high dimension of feature maps in CNNs. Thus, we calculate the WE w.r.t. each image

[1]Laplace correction ($p \longleftarrow (pN + 1)/(N + K)$) is used to avoid zero probabilities, where N is the number of training instances and K is the number of classes.

patch/region of interest (ROI) of the input images. Denoting the ROI as $\mathbf{x}_w$, we explore the importance of $\mathbf{x}_w$ by corrupting the ROI of the original image followed by analyzing the difference of prediction outcome, see Fig. 2 for further illustration. The corruption is accomplished by replacing the pixels in the ROI with samples taken directly from other images, at the same location.

Moreover, based on the following two observations in [5]: i) a pixel depends most strongly on a small neighborhood around it, ii) the conditional of a pixel given its neighborhood does not depend on the position of the pixel in the image. $p(x_i|\boldsymbol{X}_{\backslash i})$ in (2) can be approximated as below

$$p(x_i|\boldsymbol{X}_{\backslash i}) \approx p(x_i|\hat{\mathbf{x}}_{\mathbf{i}\backslash \boldsymbol{i}}) \quad (3)$$

where $\hat{\mathbf{x}}_{\mathbf{i}\backslash \boldsymbol{i}}$ is a larger region which covers x_i. The final algorithm implemented to calculate the WE w.r.t. each patch $\mathbf{x}_w$ was summarized in Algorithm 1 in [5].
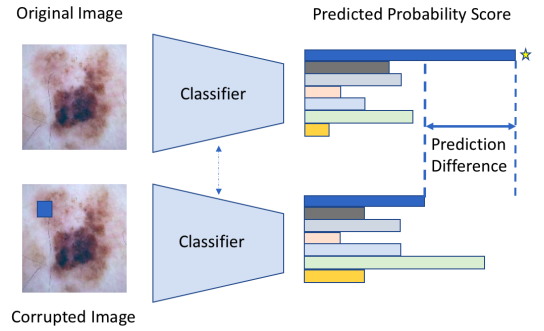

Fig. 2. Prediction difference analysis. Top: The prediction scores for each class given by the CNN classifier on the original image. The original image belongs to the class c denoted with a star. Bottom: The prediction scores for each class given by the same CNN classifier on the same image where the blue region (ROI) was corrupted. The difference in the two prediction scores illustrate the importance of the blue region in the classification decision made by the CNN classifier.

Table 1. Comparison of different model strategies

Model	VGG	ResNet	**VGG+ResNet**
Accuracy	0.79	0.82	**0.85**

Table 2. Performance summary of the ensemble model

Categories	Precision	Recall	F1 score	Samples
MEL	0.70	0.54	0.61	223
NV	0.90	0.96	0.93	1341
BCC	0.78	0.78	0.78	103
AKIEC	0.56	0.61	0.58	66
BKL	0.76	0.66	0.71	220
DF	0.83	0.65	0.73	23
VASC	0.84	0.72	0.78	29
Total	0.84	0.85	0.84	2005

3. EXPERIMENTS AND RESULTS

3.1. Dataset

The dataset used in this challenge consisted of 10015 images (327 actinic keratosis (AKIEC), 514 basal cell carcinoma (BCC), 115 dermatofibroma (DF), 1113 melanoma (MEL), 6705 nevus (NV), 1099 pigmented benign keratosis (BKL), 142 vascular lesions (VASC)) extracted from the ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection grand challenge datasets [11, 12]. Each data is RGB color image, with size 450×600.

3.2. Deep learning classification

We used 3 models: VGG16, ResNet50 and ensembled VGG16 + ResNet50 to classify the resized images. The loss function optimized to train the networks was categorical cross-entropy.

We split 70% data as training set, 10% as the validation set, which was used to find the early stopping epoch and 20% as testing set to evaluate our algorithms. The number of testing samples in each class were listed in Tabel 2.

As the number of images in each category varies widely and seriously unbalanced, we augmented the images of different classes in the training set accordingly. The augmentation methods included randomly rotation up to 25°, left-right flipping, top-bottom flipping and zoom-in cropping with ratio 0.8. All the input images were re-sized to (224,224) in our application.

The performance of each model is summarized in Table 1. We observed that the ensemble model outperformed the single model. The feature interpretation analysis described in the next subsection was applied on the ensemble model, since the more accurate the classifiers were, the more reliable the interpreted results were. The classification results of VGG16 + ResNet50 ensemble model were summarized in Table 2.

3.3. Deep learning feature interpretation

In order to interpret the features the deep learning classifier used to classify skin lesions, we did the prediction difference analysis as described in section 2.2. We investigated the ROI with length win_size in a patch with length $win\_size + pad\_size$. This patch traversed around the whole image with overlapping. Notably, each pixel of the image was visited multiple times T, except the 4 pixels on the four corners of the image. The final *weight of evidence* assigned to pixel p is $WE_p(c|\boldsymbol{X}) = \sum_t^T WE_p^t(c|\boldsymbol{X})$. We set the size of padding

to 2, which was used to find the surrounding pixels around the feature and generate the Gaussian parameters for conditional sampling. We investigated the suitable win_size to capture the predictive feature for the deep learning classifier. We set $win\_size = 5, 10, 15$ and 20. The results are shown in Fig. 3. We found that, with $win\_size = 10$ or 15, most distinguish-
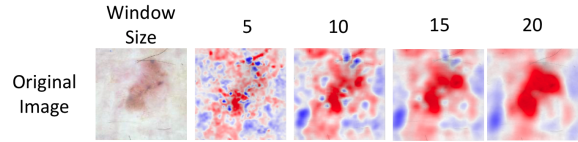
Fig. 3. Investigating different window sizes to interpret the skin lesion classification results.

ing features were captured and we could get interpretable results. Hence, we chose $win\_size = 15$ and randomly displayed the two instance of each class in Fig. 4.

3.4. Discussion

From the two instances for each class given in Fig. 4, we observed that the features contributing to the classifier (highlighted in red) followed the patterns below. For MEL, the neural network marked dark, dense, variously sized, asymmetric distributed structures [13]. For NV, the pigment network and the globular structures are marked, corresponding to the clinical evidence [14]. In addition, the boundary of the lesion was marked, where the size of the lesion might be evidence for the classifier. As for BCC, small gathered spots such as leaf like spots and blue gray dots [15] were marked. The annular-granular structures of the skin and hair follicle openings surrounded by a white halo were marked by red as evidence in AKIEC class [16]. BKL was labeled on the small nub-like extensions [17]. DF was marked on the peripheral delicate pigment network [18]. For VASC, circumscribed and ovoid structures were marked [19]. Also, it seems like the classier barely considered the surrounding skin information when recognizing VASC. The blue regions were either background or the common features shared by different classes that negatively impact the classifier.
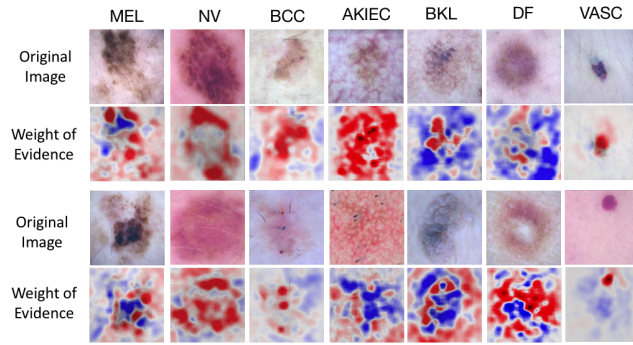
	MEL	NV	BCC	AKIEC	BKL	DF	VASC

Fig. 4. The interpretation results ($win\_size = 15$) for classifying skin lesion. Each column stands for each class. There are seven classes in our classification task. The original images are shown in the 1st and 3rd rows, where two instances of each class are given. Their weights of evidence maps are shown in the 2nd and 4th rows correspondingly. Red color highlights the evidence for the classifier and blue color highlights the evidence against the classifier.

4. CONCLUSION

In this paper, we proposed a pipeline to interpret the saliency features (biomarkers) detected by deep learning model to classify skin lesion dermoscopic images. A highly accurate deep learning classifier was trained for our investigation. From the interpreted weight of evidence maps, we found discernible features of each class. The patterns match the dermatologist criteriam identifying potential for improving clinical skin lesions detection.

5. REFERENCES

[1] Anthony F Jerant et al., "Early detection and treatment of skin cancer.," *American family physician*, vol. 62, no. 2, 2000.

[2] Michael Binder et al., "Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists," *Archives of dermatology*, vol. 131, no. 3, pp. 286–291, 1995.

[3] Lequan Yu et al., "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE transactions on medical imaging*, vol. 36, no. 4, pp. 994–1004, 2017.

[4] Xiaoxiao Li et al., "Brain biomarker interpretation in asd using deep learning and fmri," in *MICCAI*. Springer, 2018, pp. 206–214.

[5] Luisa M Zintgraf et al., "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[6] Naheed R Abbasi et al., "Early diagnosis of cutaneous melanoma: revisiting the abcd criteria," *Jama*, vol. 292, no. 22, pp. 2771–2776, 2004.

[7] Fiona M Walter et al., "Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study," *Br J Gen Pract*, vol. 63, no. 610, pp. e345–e353, 2013.

[8] "Isic 2018: Skin lesion analysis towards melanoma detection, https://challenge2018.isic-archive.com/," .

[9] Jia Deng et al., "Scalable multi-label annotation," in *ACM Conference on Human Factors in Computing Systems (CHI)*, 2014.

[10] Guolin Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *NIPS*, 2017, pp. 3146–3154.

[11] Rosendahl C.& Kittler H. Tschandl P., "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.

[12] Noel CF Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *ISBI*. IEEE, 2018, pp. 168–172.

[13] dermoscopedia Aimilios Lallas, "Superficial spreading melanoma — dermoscopedia," 2018, [Online; accessed 15-October-2018].

[14] "Classification of nevi / benign nevus pattern, https://dermoscopedia.org/classification_of_nevi_/_benign_nevus_pattern," .

[15] Natalia Jaimes Ash Marghoob, "Basal cell carcinoma — dermoscopedia," 2018, [Online; accessed 15-October-2018].

[16] Florentia Dimitriou et al., "Actinic keratosis — dermoscopedia," 2018, [Online; accessed 15-October-2018].

[17] Stephanie Nouveau dermoscopedia Ralph Braun, "Solar lentigines — dermoscopedia," 2018, [Online; accessed 15-October-2018].

[18] Pedro Zaballos dermoscopedia Ignacio Gmez Martn, "Dermatofibromas — dermoscopedia," 2018, [Online; accessed 15-October-2018].

[19] Ignacio Gmez Martn dermoscopedia Pedro Zaballos, "Vascular lesions — dermoscopedia," 2018, [Online; accessed 26-September-2018].