From Deep Learning Towards Finding Skin Lesion Biomarkers

Xiaoxiao Li[1,†], Junyan Wu[2,†,⋆], Eric Z. Chen[3,†] and Hongda Jiang[4]

***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 by eye, especially in the early stage of melanoma, is extremely challenging for the dermatologists. Hence, the discovery of reliable biomarkers will be meaningful for melanoma diagnosis. In recent years, the value of deep learning empowered computer-assisted diagnose has been shown in biomedical imaging-based decision making. However, much deep learning research focuses on improving disease detection accuracy but not understanding the features deep learning use to determine the evidence of pathology. We aim to make sure the features used by deep learning methods are the reasonable clinical features for skin lesions diagnosis, rather than artifacts. Further, we aim to discover new biomarkers, which may not have been included in clinical criteria but do make sense to the dermatologists. Our proposed pipeline can find biomarkers for identifying different lesions. The patterns are agreed with dermatologists. Surprisingly, we find surround skins also can be used as evidence for skin lesion diagnosis, which has not been included in traditional diagnosis rules. The biomarkers discovered from deep learning classifier can be significant and useful to guide clinical diagnosis.**

I. Introduction

Skin cancer is a severe public health problem in the United States, with over 5 million newly diagnosed cases every year. Melanoma, as the severest form of skin cancer, is responsible for 75% of deaths associated with skin cancer [1]. In 2015, the global incidence of melanoma was estimated to be over 350,000 cases, with almost 60,000 deaths. Fortunately, if detected early, melanoma survival exceeds 95% [2].

Dermoscopy is one of the most widely used skin imaging techniques to distinguish the lesion spots on skin [3]. It has been developed to improve the diagnostic performance of melanoma. Nevertheless, the automatic recognition of melanoma using dermoscopy images is still a challenging task due to the following reasons: the low contrast between skin lesions and normal skin regions makes it difficult to accurately segment lesion areas; the melanoma and non-melanoma lesions may have a high degree of visual similarity; 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. Some investigations attempted to apply low-level hand-crafted features to distinguish melanomas from non-melanoma skin lesions [4]. Recent works employing 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 [5]. Although these studies focused on improving computer assisted diagnostic accuracy, the diagnosis itself is hard even for experienced clinical practitioners based on dermoscopy images. The computer intervention not only assist decision making, but 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 [6], [7]. However, to the best of our knowledge, what kind of evidence deep learning models use 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 [8] and the 7-point checklist [9], etc. There is still much room to improve the understanding of melanoma recognition by reliable CNNs classifier. In this paper, we proposed a novel method based on deep convolutional neural networks and low-level image feature descriptors, which imitate clinical criteria representations, to solve skin lesion analysis towards melanoma detection problem. We aim to inspect whether the deep learning models and the dermatologists use similar criteria.

Generally speaking, our interpretation method belongs to the approach where we visualize how the network responds to a specific corrupted input image in order to explain a particular classification made by the network [7], which can more precisely locate the key features and do not need to retrain the network. Therefore, 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 analyzed the feature importance by corrupting with conditional sampling, then compare the prediction difference.

II. Method

A. Deep Learning Image Classifier

CNNs have led to breakthroughs in natural images classification and object recognition. Since CNNs have hierarchical feature learning capability, they have been widely

[1]Xiaoxiao Li, Yale University, New Haven, CT, USA.
[2]Junyan Wu, Cleerly Inc, New York City, New York, USA.
[3]Eric Z. Chen, Dana-Farber Cancer Institute, Boston, MA, USA.
[4]Hongda Jiang, East China University of Science and Technology, Shanghai, China.
[5]⋆ To whom correspondence should be addressed
[6]† These authors contributed equally.

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

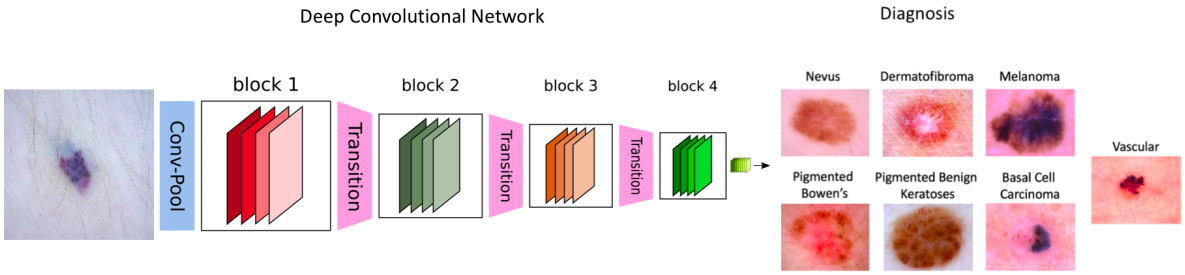

Fig. 1. The flowchart of deep learning model to classify skin lesions [2]

used in natural images classification and object recognition, due to their hierarchical feature learning capability and discrimination performance. For instance, the CNN based methods outperform the traditional techniques significantly in the recent ImageNet challenges [10]. Skin lesions have large inter classes variations. Hence the key distinguishable features that differentiate the skin lesions are difficult to be fully captured by traditional computer vision approach. We modified and fine-tuned two successful image classification CNNs architectures: ResNet50 [11] and VGG [12] to encode the image features. The fully connected(FC) layers of both networks were modified by removing the last FC layer and adding FC layer with 128 kernels and FC layer with 7 (number of classes) for the 7 class skin lesions classification task. We used LightGBM [13] to combine the different CNN model features. The LightGBM is a boosting tree-based learning algorithm. It provides multiple hyper-parameters for achieving best performance.

B. Interpreting Deep Learning Features

The interpretation method we adopted here is based on [7], which visualize how the CNNs respond to a specific corrupted input image in order to explain a particular classification made by the network. This method can localize the features contributing to the classifier. We use a heuristic method to estimate the feature (an image patch) importance by analyzing the probability of the correct class predicted by the corrupted image. Simply speaking, we corrupt the pixels in a sliding window which covers the region of interest (ROI), and then analyze the difference of prediction outcome. In this way, the relevance of features can be estimated by measuring how the prediction changes if the feature is corrupted. In the deep learning classifier case, the probability of the abnormal class c given the original image $\boldsymbol{X}$ is estimated from the predictive score of the model f : $f(\boldsymbol{X}) = p(c|\boldsymbol{X})$. We denote the image corrupted at ROI $\boldsymbol{i}$ as $\boldsymbol{X}_{\setminus \boldsymbol{i}}$. The prediction of the corrupted image is $p(c|\boldsymbol{X}_{\setminus \boldsymbol{i}})$. To calculate $p(c|\boldsymbol{X}_{\setminus \boldsymbol{i}})$, we marginalize out the corrupted ROI $\boldsymbol{i}$:

$$p(c|\boldsymbol{X}_{\setminus \boldsymbol{i}}) = \mathbb{E}_{\boldsymbol{x_i} \sim p(\boldsymbol{x_i}|\boldsymbol{X}_{\setminus \boldsymbol{i}})} p(c|\boldsymbol{X}_{\setminus \boldsymbol{i}}, \boldsymbol{x_i}), \quad (1)$$

where $\boldsymbol{x_i}$ is a sample of feature i. Once the class probability $p(c|\boldsymbol{X}_{\setminus \boldsymbol{i}})$ is estimated, we compare it with the original prediction $p(c|\boldsymbol{X})$ by calculating *Weight of Evidence (WE)*:

$$WE_i(c|\boldsymbol{X}) = log_2(odds(c|\boldsymbol{X})) - log_2(odds(c|\boldsymbol{X}_{\setminus \boldsymbol{i}})) \quad (2)$$

where $odds(c|\boldsymbol{X}) = p(c|\boldsymbol{X})/(1 - p(c|\boldsymbol{X}))$. 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. WE can be positive and negative. The feature with positive value stands for the evidence for the classifier. The feature with negative value may be the common features shared in different classes, which confuses the classifier.

Following [7], the corrupted ROI_i ($win\_size \times win\_size$) will be replaced by conditionally sampled with x_i from its surrounding neighbours. Therefore, we apply a larger patch $\hat{\mathbf{x}}_\mathbf{i}$ ($size = win\_size + pad\_size$) including feature x_i, approximating the generative model as:

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

where $p(x_i|\hat{\mathbf{x}}_{\mathbf{i} \setminus \boldsymbol{i}})$ is a multi-Gaussian model of the whole patch (window + padding). In order to examine all the feature patterns in the image, the patch was applied as a sliding window to traverse the whole image with overlapping.

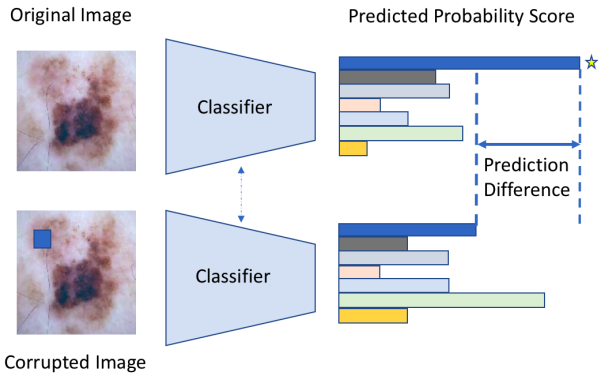

Fig. 2. Prediction difference analysis. Top: The prediction scores for each class given by the CNN classifier on the original image. The original image belongs to the class c denoted with a star. Bottom: The prediction scores for each class given by the same CNN classifier on the same image where the 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.

III. Experiments and Results

A. Dataset

The dataset was extracted from the open challenge dataset of Skin Lesion Analysis Towards Melanoma Detection (ISIC

TABLE I

COMPARISON OF DIFFERENT MODEL STRATEGIES

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

TABLE II

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

2018) [14], [15], which were collected from leading clinic centers internationally and acquired from a variety of device within each center. Broad and international participation in image contribution is designed to ensure a representative clinically relevant sample. It 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)). Each data is RGB color image, with size 450×600.

B. 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 II.

As the number of images in each category varies widely and unbalanced, we augmented the images of different classes in the training set accordingly. The augmentation methods included randomly rotation up to 25^o, 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 I. 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 II.

C. 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 II-B. 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. For visualization, we colored pixels with *positive WE* in red and pixels with *negative WE* in blue. 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 distinguishing features were captured and showed interpretable results. Hence, we chose $win\_size = 15$ and randomly displayed the two instances of each class in Fig. 4.

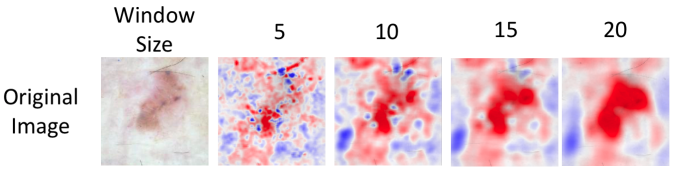

Fig. 3. Investigating different window sizes to interpret the skin lesion classification results.

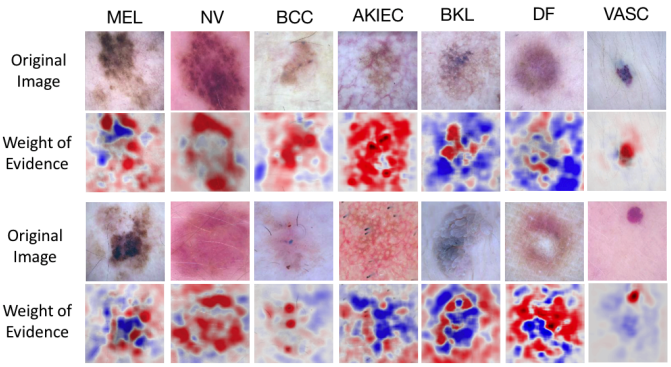

Fig. 4. The interpretation results ($win\_size = 15$) for classifying skin lesion. Each column stands for each class. There are seven classes in our classification task. The original images are shown in the 1^{st} and 3^{rd} rows, where two instances of each class are given. Their weights of evidence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

D. 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. Further, we validated the results and asked for opinions from 2 dermatologists in the different institutions. For MEL, the neural network marked dark, dense, variously sized, asymmetric distributed structures, as validated by dermatologists that Blue white veils, dark brown homogeneous areas that lack pigment network were picked which are the clinical

evidence for MEL, suggested by the dermatologists. For NV, the pigment network and the globular structure are marked, corresponding to the clinical evidence [16] and validated by the dermatologists. 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 were marked, which are blue gray globular pointed out by the dermatologists. In AKIEC class, the dermatologists pointed out the annulargranular structures of the skin and hair follicle openings surrounded by a white halo were marked by red, which are clinical evidence for AKIEC class [17]. Small nub-like extensions were highlighted in BKL[18]. The dermatologists also comfirmed that DF was marked on the peripheral delicate pigment network [19]. For VASC, circumscribed and ovoid structures [20], which were thought as lucunae of the blood vessels by the dermatologists were marked. Also, it seems like the classifier barely considered the surrounding skin information when recognizing VASC. It is interesting to see that surrounding skins can be used as evidence to classify skin lesions in CNNs. The blue regions were either background or the common features shared by different classes that negatively impact the classifier.

IV. 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. Our ensemble CNNs classifier conducted on the open challenge dataset of Skin Lesion Analysis Towards Melanoma Detection (ISIC 2018) demonstrated the effectiveness of the proposed method, even with limited and unbalanced training data. From the interpreted weight of evidence maps, we found discernible features of each class. The patterns match the dermatologist criteria identifying potential for improving clinical skin lesions detection. Further work will improve the classifier's performance, including investigating integrating probability map generated by segmentation in our networks, will try more interpretation methods (such as LIME [21], DeepLift [22], and Integrated Gradients [23]) and will explore more applications in different disease .

References

[1] Anthony F Jerant, Jennifer T Johnson, Catherine Demastes Sheridan, and Timothy J Caffrey, "Early detection and treatment of skin cancer.," *American family physician*, vol. 62, no. 2, 2000.

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

[3] Michael Binder, Margot Schwarz, Alexander Winkler, Andreas Steiner, Alexandra Kaider, Klaus Wolff, and Hubert Pehamberger, "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.

[4] Jufeng Yang, Xiaoxiao Sun, Jie Liang, and Paul L Rosin, "Clinical skin lesion diagnosis using representations inspired by dermatologist criteria," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1258–1266.

[5] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE transactions on medical imaging*, vol. 36, no. 4, pp. 994–1004, 2017.

[6] Xiaoxiao Li, Nicha C Dvornek, Juntang Zhuang, Pamela Ventola, and James S Duncan, "Brain biomarker interpretation in asd using deep learning and fmri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 206–214.

[7] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[8] Naheed R Abbasi, Helen M Shaw, Darrell S Rigel, Robert J Friedman, William H McCarthy, Iman Osman, Alfred W Kopf, and David Polsky, "Early diagnosis of cutaneous melanoma: revisiting the abcd criteria," *Jama*, vol. 292, no. 22, pp. 2771–2776, 2004.

[9] Fiona M Walter, A Toby Prevost, Joana Vasconcelos, Per N Hall, Nigel P Burrows, Helen C Morris, Ann Louise Kinmonth, and Jon D Emery, "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.

[10] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei, "Scalable multi-label annotation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 3099–3102.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[14] 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.

[15] 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.

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

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

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

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

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

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[22] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.

[23] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.