Computer-Aided Clinical Skin Disease Diagnosis Using CNN and Object Detection Models

Xin He*, Shihao Wang*, Shaohuai Shi*, Zhenheng Tang*,
Yuxin Wang*, Zhihao Zhao*, Jing Dai*, Ronghao Ni*,
Xiaofeng Zhang†, Xiaoming Liu‡, Zhili Wu§, Wu Yu§, Xiaowen Chu*¶
*Department of Computer Science, Hong Kong Baptist University
†Harbin Institute of Technology, China
‡The University of Hong Kong - Shenzhen Hospital, China
§Shenzhen Yiyuan Intelligent Technology Co., Ltd., China

***Abstract*—Skin disease is one of the most common types of human diseases, which may happen to everyone regardless of age, gender or race. Due to the high visual diversity, human diagnosis highly relies on personal experience; and there is a serious shortage of experienced dermatologists in many countries. To alleviate this problem, computer-aided diagnosis with state-of-the-art (SOTA) machine learning techniques would be a promising solution. In this paper, we aim at understanding the performance of convolutional neural network (CNN) based approaches. We first build two versions of skin disease datasets from Internet images: (a) Skin-10, which contains 10 common classes of skin disease with a total of 10,218 images; (b) Skin-100, which is a larger dataset that consists of 19,807 images of 100 skin disease classes. Based on these datasets, we benchmark several SOTA CNN models and show that the accuracy of skin-100 is much lower than the accuracy of skin-10. We then implement an ensemble method based on several CNN models and achieve the best accuracy of 79.01% for Skin-10 and 53.54% for Skin-100. We also present an object detection based approach by introducing bounding boxes into the Skin-10 dataset. Our results show that object detection can help improve the accuracy of some skin disease classes.**

***Index Terms*—computer-aided skin disease diagnosis, CNN, ensemble method, object detection**

I. Introduction

It is generally known that there are many types of skin diseases, ranging from itching caused by mosquito bites to skin cancer. Traditionally, the diagnosis of skin diseases is based on the comprehensive consideration of the size, shape, color and other visual features of the lesion area. Recently, deep learning methods have been applied to many medical tasks [1], [2] and obtained remarkable achievements. Using computers to help diagnose skin diseases would be a promising direction. At present, there are many studies [3], [4] on dermoscopic images and they have achieved promising results. Although there are some works [5], [6] on clinical skin disease images, their datasets are small. To fill this gap, we first build a clinical skin disease dataset, namely Skin-10, which has 10,218 images of 10 common skin diseases, and we manually marked the skin lesions for each image. Besides, we expand Skin-10 to Skin-100, which covers 100 classes of skin diseases and contains 19,807 clinical images. To the best of our knowledge, the scale of Skin-100 is larger than all existing clinical skin disease datasets. What's more, Skin-10 is the first clinical skin disease dataset that provides bounding boxes. By using the bounding boxes, we could extract more discriminative features and thereby improve the classification performance.

Unlike dermoscopic images, which are hard to obtain because of high expense and inconvenience access [10], clinical images can be captured by easily-accessed devices, like the smartphone. However, there are several difficulties in recognizing skin lesion from a clinical image. (a) the background of clinical images is more complex so that how to reduce background noise interference is an important issue to consider. (b) clinical images cover far more classes of diseases than dermoscopic images, because dermatoscope is designed primarily for skin cancers, like melanoma and basal cell carcinoma, etc. (c) clinical skin disease image classification can be considered as a task of fine-grained classification, which is more difficult than normal classification task (e.g. CIFAR10 [11] and ImageNet [12]). On the one hand, the lesions of one class of skin diseases may not only appear in different parts of the human body, but also the visual characteristics of these lesions vary greatly, which indicates the high intra-class variance. On the other hand, two lesions that look very similar may belong to different categories, which indicates the low inter-class variance. Despite the above difficulties, it would be helpful for doctors and patients if we can apply deep learning techniques on our datasets to solve those problems.

To verify whether deep neural network works well on the task of clinical skin disease classification, we first benchmark on Skin-10 and Skin-100 using four SOTA CNNs: ResNet50 [13], DenseNet121 [14], Nasnetamobile [15] and Pnasnet5large [16], and we implement ensemble methods based on above four CNN models. Besides, we perform two SOTA object detection models (RetinaNet [17] and Faster-RCNN [18]) on Skin-10 to detect possible lesion regions on

¶Corresponding author: chxw@comp.hkbu.edu.hk

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

TABLE I
THE SUMMARY OF EXISTING SKIN DISEASE DATASETS. IN THE FIRST ROW (TYPE), D AND C INDICATES DERMOSCOPIC AND CLINICAL, RESPECTIVELY. THE FOURTH ROW (ANNOTATION?) REPRESENTS WHETHER THE DATASET PROVIDES BOUNDING BOXES OR SEGMENTATION.

Dataset	PH2 [3]	ISIC	Ham10000 [4]	[7]	[5]-1^{st} / -2^{nd}	[8]	[6]-1^{st} / -2^{nd}	[9]	Skin-10 / Skin-100
Type	D	D	D	C	C / C	C	C / C	C	C / C
#Classes	3	-	7	6	3 / 7	44	128 / 198	26	10 / 100
#Images	200	23,801	10,015	366	90 / 706	2,309	5,619 / 6,584	17,777	10,218 / 19,807
Annotation?	Y	Y	N	N	N / N	N	N / N	N	Y / N
Year	2013	2016	2018	1998	2012	2013	2016	2019	2019

the raw images and then classify the disease based on the region with the highest confidence score. The experimental result shows that the object detection based method can reduce the influence of the background and achieve higher accuracy than plain CNN models.

In summary, this paper has the following contributions. (a) We build two versions of clinical skin disease datasets: Skin-10 and Skin-100. As far as we know, Skin-10 is the first clinical skin disease dataset providing bounding boxes and the scale of Skin-100 is larger than most existing clinical skin disease datasets. (b) We establish the baseline performance of four SOTA CNN models on our datasets. (c) We verify the effectiveness of the ensemble method which achieves the best accuracy on both datasets. (d) We propose an object detection based approach and evaluate its performance on Skin-10.

The remainder of this paper is organized as follows. Section II reviews the existing related skin disease datasets and methods of recognizing skin diseases. In Section III, the details of Skin-10 and Skin-100 are presented. Section IV describes the details of CNN-based, ensemble, and object detection based methods, and analyzes the experimental results. Finally, Section V offers concluding remarks and future works.

II. RELATED WORK

A. Skin Disease Datasets

There are two types of skin disease images: dermoscopic and clinical. The former is obtained by a dermatoscope, which requires a high-quality magnifying lens and a powerful lighting system. Hence, dermoscopic images have a simpler background, which means that the distribution, number, size, shape, and color of skin lesions are much clearer compared with clinical images. Table I summarizes some existing skin disease datasets. PH2 dataset [3] contains 200 dermoscopic images of melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 melanomas. HAM10000 [4] collects dermoscopic images from different populations and consists of 10,015 images from 7 classes. ISIC Archive[1] provides both dermoscopic and clinical images, which are categorized into many classes based on different attributes, such as diagnostic attributes, clinical attributes, etc. ISIC has 23,801 dermoscopic images but only has 100 clinical skin images. The dataset [7] is from the UCI machine repository and contains six classes of skin diseases. In [5], there are two datasets proposed: 90 dermatological images covering 3 skin disease classes and 706 images covering 7 classes, respectively. The dataset in [8] has 44 classes containing 2,309 images. Another two datasets are proposed in [6]: SD-128, which has 128 classes and 5,619 images, and SD-198, which has 198 classes and 6,584 images. They achieve the classification accuracy of 52.15% on SD-128 and 50.27% on SD-198. A recent work proposed by Google [9] develops a deep learning system for diagnosing 26 classes of clinical skin diseases using 17,777 cases. Each case consists of one or several clinical images and metadata, which includes patient demographic information and medical history.

B. Skin Disease Classification

The methods of skin disease image classification are two-folds. The first relies on hand-crafted features, such as the texture (SIFT, HOG), color (ColorSIFT, ColorHistogram), and edge (Gabor, Sobel) [6]. The models, like support vector machines (SVMs), k-nearest neighbors (KNN) and decision trees, are then trained based on above features. The second is deep learning techniques. In [6], a pretrained CNN model is used to extract deep features and the result shows that CNN performs better than hand-crafted based methods when the background is complex. In [19], the authors use a large ensemble of SOTA CNN models and place second at the ISIC 2018 challenge for skin lesion diagnosis. Verma and Pal et al. [20] also propose an ensemble method that combines five different data mining techniques. Zhang's work [21] proposes an attention residual learning CNN model to effectively locate the skin lesion of dermoscopic images. Currently, the models generated by the neural architecture search (NAS) technique have been demonstrated to achieve comparable results to the human-designed models in many tasks [22], [23]. In this paper, we compare NAS-designed and human-designed CNN models and perform the ensemble method based on these two types of models. We also evaluate the performance of two SOTA object detection models, which are expected to locate the skin lesion area and reduce the influence of background.

III. DATASET

Our datasets are built by scraping images from the Internet. We build Skin-10 by selecting 10 classes from the most common skin diseases [24]. As a result, Skin-10 has 10,218 images and the statistics of Skin-10 is presented in Table II. Additionally, we use a graphical image annotation tool (LabelImg [2]) to mark skin lesions with bounding boxes for each image in Skin-10. We further build a larger dataset based

[1]https://isic-archive.com/

[2]https://github.com/tzutalin/labelImg

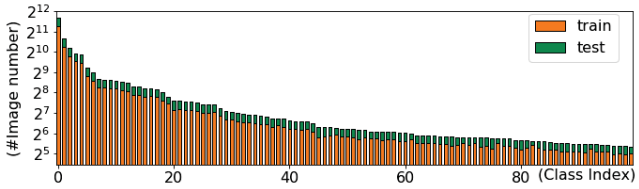

Fig. 1. Statistics of Skin-100. A base-2 log scale is used for the Y-axis. For each class, the number of images in the training set is about 3 times the testing set.

TABLE II
THE STATISTICS OF SKIN10. FOR NOTATIONAL CONVENIENCE, WE WILL USE THE INDEX VALUE TO INDICATE THE CORRESPONDING DISEASE CATEGORY IN THE FOLLOWING CONTENT UNLESS OTHERWISE STATED.

Index	Class name	#Training set	#Testing set
0	Acne Vulgaris	1598	399
1	Actinic Keratosis	932	239
2	Atopic Dermatitis Eczema	590	145
3	Basal Cell Carcinoma	3249	826
4	Compound Nevus	513	127
5	Onychomycosis	394	98
6	Rosacea	976	242
7	Seborrheic Keratosis	1180	291
8	Stasis Ulcer	407	100
9	Tinea Corporis	379	87

on Skin-10, namely Skin-100, which has 19,807 images of 100 skin diseases, and each class has over 40 images. The data distribution of Skin-100 is long-tailed, as shown in Fig 1. In both Skin-10 and Skin-100, the ratio of the training set to the testing set is set at 3:1.

In some images, the skin lesion is covered or cured. Hence, we perform data cleaning to remove a total of 290 noise images. The experimental result in Section 4 shows that data cleaning improves the performance of models for all CNN models.

Scale To the best of our knowledge, the scale of Skin-100 is much larger than existing clinical skin disease datasets. As Table I shows, the number of clinical skin disease images in Skin-100 is almost 3 to 200 times than other datasets. Besides, Skin-10 is the first clinical skin disease dataset providing bounding boxes for skin lesion detection [5], [6], [8].

Diversity Our datasets cover different ages, genders and lesion locations. Besides, the difference can be significant within the same class, e.g. the skin disease images from the same class may differ from skin colors and lesion shapes, whereas, the difference can also be subtle between different classes. In a word, our datasets are of high diversity so that it is worthy to evaluate whether deep leaning techniques are feasible in our datasets.

IV. CLINICAL SKIN DISEASE IMAGE CLASSIFICATION

In order to establish the baseline performance of CNN models on our datasets and not lose generality, we select four representative models from two types of models. (a) the models designed by human experts. (b) the models generated by the NAS algorithm. We also verify the feasibility of the ensemble method using these four models. We further evaluate the effectiveness of two SOTA object detection models. The implementation details and results are described in the following content.

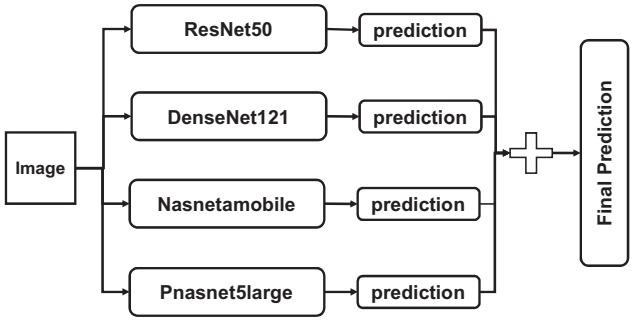

Fig. 2. An overview of EnsembleNet.

A. CNN based Classification

In this experiment, two types of SOTA CNN models are used as the baseline models: (a) ResNet50 and DenseNet121, which are designed by human experts. (b) Nasnetamobile and Pnasnet5large, which are generated automatically by the NAS technique. The pretrained models of (a) and (b) are obtained from torchvision[3] and pretrained-models.pytorch[4], respectively. After fine-tuning four base CNN models, we ensemble them into a strong classifier, namely EnsembleNet (shown in Fig 2), by summing the probability prediction of four base models:

$$\text{EnsembleNet}(x) = \sum_{k}^{K} \text{BaseNet}_k(x) \tag{1}$$

where x indicates the input image and K represents the number of base models. In this work, the best result is obtained by setting $K = 4$, i.e. combining the prediction results of all base models.

1) Settings: Before feeding the training set to the model, we implement a series of data augmentation, including resize, random crop, flips, rotation, and normalization, while for the testing set, we only perform resize and normalization. We use the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01. The learning rate will be multiplied by a factor of 0.1 every 10 epochs. The batch size is 64 and the input image size is fixed to 224*224. Cross-entropy is used as the loss function. All baseline models are fine-tuned to converge.

2) Results and Analysis: Based on the observation that the distribution of our datasets is imbalanced, we use top-k weighted accuracy as the metric. Let $\hat{y}_i$ be the output of the CNN models when the input is x_i and T_i^k be the labels of the k largest elements in $\hat{y}_i$. y_i is the ground-truth label of the

[3]https://github.com/pytorch/vision/tree/master/torchvision/models
[4]https://github.com/Cadene/pretrained-models.pytorch

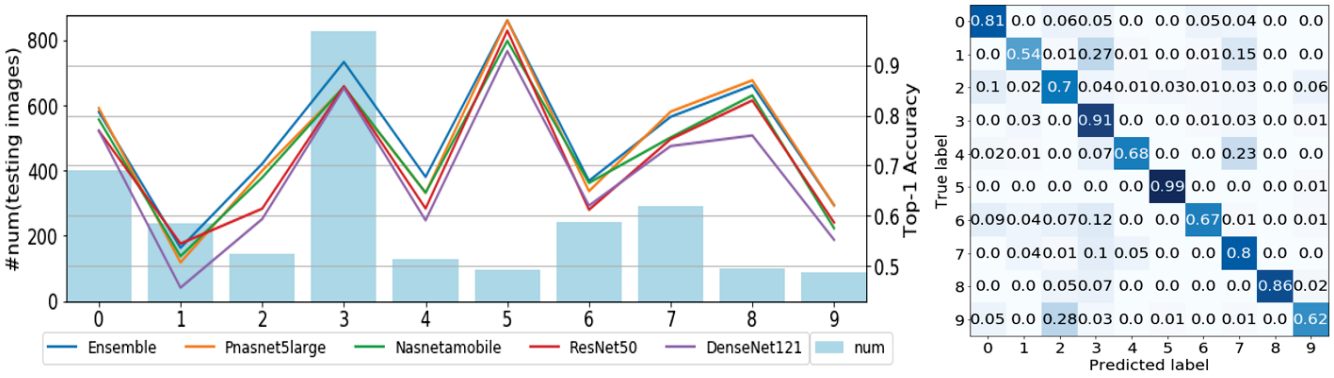

Fig. 3. Left: The results of different models on Skin-10. Right: the confusion matrix of EnsembleNet. The number 0-9 in the x-axis indicates different class index.

TABLE III
THE CLASSIFICATION ACCURACY OF CNN MODELS ON OUR DATASETS.

Dataset	Skin-10		Skin-100		Skin-100-Noise	
Top-k Accuracy (%)	Top-1	Top-3	Top-1	Top-5	Top-1	Top-5
ResNet50	74.75	94.28	47.70	74.94	40.36	68.18
DenseNet121	72.94	92.68	48.43	75.37	45.08	70.76
Nasenetamobile	75.8	94.01	46.69	73.57	45.53	73.29
Pnasnet5large	76.94	94.36	48.24	73.66	47.24	73.32
EnsembleNet	**79.01**	**95.34**	**53.54**	**79.18**	**51.96**	**78.51**

i-th sample and w_i is the corresponding sample weight. The top-k weighted accuracy on the testing set is given as

$$A_{\text{top}-k} = \sum_{i=1}^{N} Z_i^k w_i \tag{2}$$

where

$$w_i = \frac{\sum_j 1\left(y_j = y_i\right)}{N}$$
$$Z_i^k = \begin{cases} 1, & y_i \in \mathbf{T}_i^k \\ 0, & \text{otherwise} \end{cases}$$

and N is the number of images in the test set.

Table III presents the results of the baseline models on Skin-10, Skin-100 and Skin-100-Noise. Skin-100-Noise represents the dataset in which there exist many noise images in the training set, while Skin-100 is the opposite. Additionally, we make sure that both Skin-100 and Skin-100-Noise share the same testing set. In this way, we can fairly see from Table III that removing the noise images contributes to the improvement of classification accuracy for all CNN models, especially for ResNet50, as its top-1 accuracy is improved by 7.34%. On the other hand, we can also find that, after data cleaning, the performance of models designed by the NAS technique is not improved as much as the human-designed models, which means that these models are more robust and able to extract more discriminative features.

We further employ an ensemble method and we evaluate all possible combinations of base models. The experimental results show that the ensemble of four base models (EnsembleNet) achieves the highest top-1 accuracy (79.01%), and the ensemble of DenseNet121, Nasnetamobile, and Pnasnet5large obtains the best top-3 accuracy (95.42%). What's more, all ensemble models outperform any single base model.

Although EnsembleNet can improve classification accuracy by integrating base models, we can see from Fig 3 (left) that EnsembleNet and base CNN models have similar performance on Skin-10, e.g. they all achieve high score on class 3, 5 and 8, but perform not well on class 1, 2, 4, and 9. Fig 3 (right) plots the confusion matrix of EnsembleNet, from which we can see that the classes that confuses EnsembleNet can be divided into two groups: (a) class 0, 2 and 9. (b) class 1, 3, 4, 6 and 7. The reason why the accuracy of class 0 and 3 is relatively higher probably is that they have much more images. But why is the accuracy of class 5 and 8 is high even though the number of images in both classes is not as many as class 0 and 3? Based on the observation of the images of different classes, we find that most of the lesion locations of class 5 and 8 are nails and legs, respectively, while the lesions of other classes are distributed in similar regions (e.g. faces, chest or unknown area) and their visual features also look similar. Hence, we can conclude that for class 5 and 8, most of their lesions are distributed in the specific regions, so the CNN models can correctly classify them even if it does not correctly locate the lesion regions. However, for some classes with similar lesion locations, the CNN models may misclassify the images due to focusing on the wrong region or failure to extract effective and discriminative features from the lesion area.

B. *Object Detection based Classification*

To alleviate the problem of the plain CNN models, we evaluate the performance of two SOTA object detection models (RetinaNet and Faster-RCNN), respectively, which are expected to learn to detect and classify based on the skin lesion region. The process is shown in Fig 4.

1) Settings: The implementation of both RetinaNet and Faster-RCNN are publicly available in mmdetction[5]. The

[5]https://github.com/open-mmlab/mmdetection

TABLE IV
THE RESULTS OF DIFFERENT MODELS ON SKIN-10. THE INDEX 0 TO 9 INDICATE THE CLASSIFICATION ACCURACY OF DIFFERENT CLASSES. THE LAST COLUMN IS THE OVERALL WEIGHTED ACCURACY ON SKIN-10.

Model	Top-1 Accuracy(%)										
	0	1	2	3	4	5	6	7	8	9	Overall
ResNet50	76.94	54.39	61.38	85.84	61.42	96.94	61.16	75.26	83.00	58.62	74.75
DenseNet121	76.94	45.61	59.31	85.47	59.06	92.86	61.98	73.88	76.00	55.17	72.94
Nasnetamobile	79.20	51.88	67.59	85.71	64.57	94.90	66.53	75.60	84.00	57.47	75.80
Pnasnetamobile	**81.45**	50.63	68.97	85.59	64.57	**98.98**	64.88	80.76	**87.00**	62.07	76.94
Faster-RCNN	79.70	60.67	47.59	87.17	**67.72**	**98.98**	**77.69**	78.69	76.00	63.22	77.64
RetinaNet	80.95	**61.09**	**68.97**	**88.01**	55.91	96.94	69.10	**83.51**	71.0	**65.52**	**78.31**

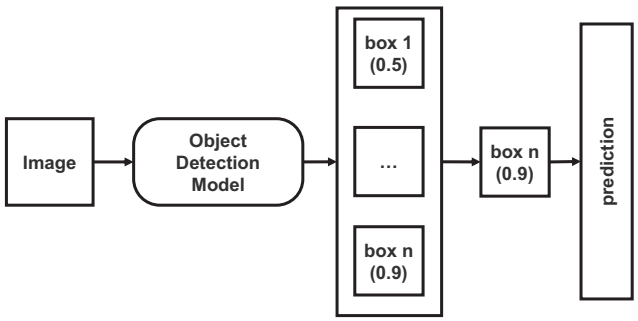

Fig. 4. An overview of object detection based method. The final prediction is made based on the box with the highest confidence score.

backbone of both models is ResNeXt101_64x4d. The input images are zoomed to 400×400. The model is optimized by SGD with an initial learning rate of 0.001, and the learning rate is dropped by a factor of 10 every 10 epochs. Besides, for RetinaNet, the focal loss [17] is used to replace the commonly used cross-entropy loss, and we set $\alpha = 0.25, \gamma = 2$.

2) Results and Analysis: Mean Average Precision (mAP) is one of the most commonly used metric for the object detection task. However, for Skin-10, each image corresponds to only one class of skin diseases, and we care more about whether the image is classified correctly. Therefore, we take the prediction of the box, which has the highest confidence, as the prediction of the image.

Let $\hat{y}_i$, y_i be the prediction, ground truth for the i-th image, respectively. w_i is the sample weight for i-th image, as defined in Equation 3. N and C indicate the number of testing images and classes, respectively. Then the top-1 classification accuracy is defined as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{\sum_i^N \mathbb{1}\left(\hat{y}_i, y_i\right)}{N}, \mathbb{1}(p,q) = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases} \\ \text{s.t. } \hat{y}_i &= \arg\max_n b_i^{mn}, m \in [1, M_i], n \in [1, C] \end{aligned} \tag{3}$$

where M^i represents the number of predicted boxes in the i-th image. b_i^{mn} indicates the probability of being predicted as the n-th class for the m-th box in the i-th image.

It can be seen from the result in Table IV that the classification accuracy of most classes is improved by detecting skin lesions. However, we can also find that the classification accuracy of some classes (e.g. class 8) decreases significantly for both RetinaNet and Faster-RCNN, instead. There are several possible explanations for this result. First, the prediction is only based on the box with the highest confidence. In other words, it loses the global features that are important for classification, which further confirms the conclusion of the previous experiment that the reason why base CNN models can achieve higher classification accuracy in some classes (e.g. class 8) is because lesion locations of these classes are unique and not easy to be confused by other classes. Second, although both object detection models can detect lesions well in most cases, it may detect the wrong area or even fail to detect any targets when the skin lesion area is not clear or affected by background noise, thus leading to failure of prediction.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated how deep learning techniques could help image-based skin disease diagnosis. We developed two versions of clinical skin disease datasets from Internet images: Skin-10, which consists of 10,218 images of 10 common skin disease classes with bounding boxes surrounding the lesion, and Skin-100, which contains 19,807 images of 100 skin disease classes. We found that data cleaning is very important as it can help improve the top-1 accuracy by 4% on average in our experiments. We also found that the ensemble method makes more efficient use of the dataset and outperforms any single CNN model on both Skin-10 and Skin-100. We further evaluated two SOTA object detection models that are used to reduce the influence of image background. Our results showed that the object detection models outperform the classification based solutions.

Although it is demonstrated that deep learning techniques can achieve satisfactory performance in skin disease image classification, there still exists room for further improvement. Based on the analysis of previous experimental results, we summarize the following directions worth studying in the future. First, the models generated by autoML technique may extract more discriminative features and are more robust to noisy data than human-designed models, therefore we can try to explore autoML technique to generate a task-specific architecture based on clinical skin disease datasets. Second, although the ensemble method is easy to implement and effective, it does not improve the accuracy of some hard-to-classify classes. Thus, how to improve the accuracy of these classes is still a challenging task. Third, object detection based

approach can reduce the influence of background by detecting the local region of skin lesions, but it loses the global features. Hence, how to combine global and local features would be another promising work. At last, our dataset is built on Internet images and cannot be open to the public due to copyright issue. We hope the healthcare community can join hands to develop an open and large skin disease dataset.

REFERENCES

[1] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, 2017.

[2] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[3] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 2013, pp. 5437–5440.

[4] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *arXiv preprint arXiv:1803.10417*, 2018.

[5] O. Razeghi, G. Qiu, H. Williams, K. Thomas, and I. VIPLAB, "Skin lesion image recognition with computer vision and human in the loop," *Medical Image Understanding and Analysis (MIUA), Swansea, UK*, pp. 167–172, 2012.

[6] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *European Conference on Computer Vision*. Springer, 2016, pp. 206–222.

[7] H. A. Güvenir, G. Demiröz, and N. Ilter, "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals," *Artificial Intelligence in Medicine*, vol. 13, no. 3, pp. 147–165, 1998.

[8] O. Razeghi, Q. Zhang, and G. Qiu, "Interactive skin condition recognition," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2013, pp. 1–6.

[9] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. d. O. Marinho, J. Gallegos, S. Gabriele *et al.*, "A deep learning system for differential diagnosis of skin diseases," *arXiv preprint arXiv:1909.05382*, 2019.

[10] M. W. Tsang and C. L. Kovarik, "Global access to dermatopathology services: physician survey of availability and needs in sub-saharan africa," *Journal of the American Academy of Dermatology*, vol. 63, no. 2, pp. 346–348, 2010.

[11] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[15] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 8697–8710.

[16] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 19–34.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[19] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, "Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting," *arXiv preprint arXiv:1808.01694*, 2018.

[20] A. K. Verma, S. Pal, and S. Kumar, "Classification of skin disease using ensemble data mining techniques," *Asian Pacific Journal of Cancer Prevention*, vol. 20, no. 6, pp. 1887–1894, 2019.

[21] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Transactions on Medical Imaging*, 2019.

[22] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *arXiv preprint arXiv:1908.00709*, 2019.

[23] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *arXiv preprint arXiv:1808.05377*, 2018.

[24] R. J. Hay, N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf *et al.*, "The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions," *Journal of Investigative Dermatology*, vol. 134, no. 6, pp. 1527–1534, 2014.