XiangyaDerm: A Clinical Image Dataset of Asian Race for Skin Disease Aided Diagnosis

Bin Xie[1,3], Xiaoyu He[1], Shuang Zhao[2,3], Yi Li[1,3], Juan Su[2], Xinyu Zhao[1], Yehong Kuang[2], Yong Wang[1,3], and Xiang Chen[2]

[1] School of Automation, Central South University, Changsha, China
{xiebin, hexiaoyu, shuangxy}@csu.edu.cn
[2] Department of Dermatology, Xiangya Hospital Central South University, Changsha, China
[3] Mobile Health Ministry of Education - China Mobile Joint Laboratory, Central South University, Changsha, China
chenxiangck@126.com

Abstract. Skin disease is a quite common disease of human beings, which has been found in all races and ages. It seriously affects people's quality of life or even endangers people's lives. In this paper, we propose a large-scale, Asian-dominated dataset of skin diseases with bounding box labels, namely XiangyaDerm. It contains 107,565 clinical images, covering 541 types of skin diseases. Each image in this dataset is labeled by professional doctors. As far as we know, this dataset is the largest clinical image dataset of Asian skin diseases used in Computer Aided Diagnosis (CAD) system worldwide. We compare the classification results of several advanced Convolutional Neural Networks (CNNs) on this dataset. InceptionResNetV2 is the best one for 80 skin disease classification whose Top-1 and Top-3 accuracies can reach 0.588 and 0.764, which proves the usefulness of the proposed benchmark dataset, and gives the baseline performance on it. The cross-test experiment with Derm101 shows us that the CNN model has a very different test effect on different ethnic datasets. Therefore, to build a skin disease CAD system with high performance and stability, we recommend to establish a specific dataset of skin diseases for different regions and races.

Keywords: skin disease, clinical image dataset, computer aided diagnosis.

1 Introduction

Skin disease is a very common disease of human beings, which has been found in all races and ages [1]. Skin diseases can bring many troubles to patients, such as itching, bleeding and so on, which seriously affect people's quality of life or even endanger lives. Early diagnosis of skin diseases is very important, which can make patients get correct treatment as soon as possible and arrest the growth of the disease. However, due to the limited medical knowledge of patients and the disparity of medical resources, such case of delaying the timing of diagnosis occurs from time to time. The emergence of the CAD system can help us solve these problems to a certain extent.

Early studies on skin diseases CAD system are mostly focused on dermoscopic images [2, 3]. This is because they focus more on lesions than clinical images with uniform illumination and less noise. In fact, dermoscopic-based diagnosis of skin diseases has some limitations in promotion, such as high fees and less convenience. In recent years, some researchers begin to pay more attention to clinical images [4, 5, 6, 7]. Their works' datasets were mainly collected in Europe and America. The lack of a specific Asian skin disease dataset has become a major hindrance to the study of skin disease diagnostic system.

Convolution neural network is very popular in the field of feature learning and object recognition in recent years. Many studies from ImageNet's large-scale visual challenge [8, 9, 10, 11] (ILSVRC) [12] show that the most advanced CNN has exceeded human level in object classification tasks. However, the classification performance of CNN sometimes depends too much on the dataset. We designed cross-test experiments to study this problem.

The contributions of this paper are summarized below. Firstly, a large-scale, Asian-dominated dataset of skin diseases is proposed. Secondly, in order to evaluate the usefulness of our dataset, we give the baseline performance on this dataset. Finally, through cross-test experiments between different datasets, we draw the conclusion that the skin disease diagnosis systems should be setup on specific datasets. We have good reason to believe that the dataset proposed in this paper is very urgent and meaningful for the research of skin disease diagnosis.

2 Related work

Esteva et al. [13] achieved good recognition rate between keratinocytic carcinoma and benign seborrheic keratosis, malignant melanoma and benign nevus using InceptionV3 CNN architecture on Dermofit and ISIC datasets., reaching the level of human dermatologists. This landmark research has attracted wide attention, especially in the field of AI in skin diseases.

Sun et al. [4] introduced datasets SD-198 and SD-128 based on DermQuest (now DermQuest is merged into Derm101). Several kinds of manual features extraction methods and deep learning methods are compared on these two datasets. SD-198 contains 198 different diseases, a total of 6,584 images. SD-128 is a subset of SD-198, ensuring that each class has more than 20 images. This benchmark dataset encourages many studies about visual skin disease classification. However, they classify 198 or 128 categories of skin diseases using a dataset of 6,584 images, which seems too small for CNN because the average number of training sets and test sets for per category is only 50.

Liao et al. [7] collected their dataset from 6 public dermatology atlas websites: AtlasDerm, Danderm, Derma, DermIS, Dermnet and DermQuest. They use CNNs for disease-targeted and lesion-targeted classifications and draw a conclusion that the classification method with lesion tags can get better performance. Their work is very meaningful both in methods and datasets. Next, we will briefly introduce the datasets mentioned above.

Dermofit dataset is provided by researchers at the University of Edinburgh in the United Kingdom. This dataset is of high quality and widely used by researchers, but it is not free available. Dermofit includes 10 types of skin diseases: actinic keratosis, basal cell carcinoma, melanocytic nevus, seborrheic keratosis, squamous cell carcinoma, intraepithelial carcinoma, pyogenic granuloma, hemangioma, dermatofibroma and malignant melanoma., but the total number of images is only about 1,300.

ISIC dataset comes from the International Skin Imaging Collaboration (ISIC), which aims to promote the diagnostic ability of skin image data. The dataset contains 23,906 images of 16 types of skin diseases, including both dermoscopic images and clinical images, with high quality and no watermarks. Each image in this dataset contains the tags of patient's age, gender, and lesion size. However, there are only 100 clinical images in the dataset, including 37 melanomas, 40 basal cell carcinomas and 23 squamous cell carcinomas, which is too small for the training of deep learning methods.

Derm101 is a website for providing clinicians with high-caliber and up-to-date content. It also provides a clinical dataset of 22,979 images of 525 types of skin diseases, containing labels both for disease diagnosis and lesion location, without watermarks on the image too. Fortunately, we have obtained permission from the Derm101 team to use their images for research purposes. Later, we organized experiments related to this dataset.

Dermnet is called to be the largest independent photo dermatology source dedicated to online medical education. The image library of Dermnet is nearly 18,974 images, 626 types of skin diseases. However, each image in this dataset has only the label of the disease diagnosis, without any other labels.

DermIS is a free dataset website built by the University of Heidelberg, Germany. There are 7,172 images in this database, which are divided into 735 categories. Each image in this dataset has regular disease diagnosis labels as well as the text descriptions of lesion location, race and age. The drawback of this dataset is that the number of images in each class is not large and the image is watermarked.

AtlasDerm is a Brazilian dataset website. There are 9,503 images and 534 categories of skin diseases. Most of the data are mainly about Brazilians in South America. Each image in this dataset has only the label of disease diagnosis, and the image is watermarked.

Danderm is a clinical image data collection website of skin diseases from Denmark. There are 1,110 images and 91 types of skin diseases. Most of the patients collected in this dataset are white races, only have the label of disease diagnosis, and contain the watermarking occlusion.

In a conclusion, we summarize the above datasets into Table 1. It can be seen from Table 1 that there are some obvious defects in the existing skin disease datasets:

1) The current datasets are mainly based on Caucasian and Black races in Europe and America, and the large-scale standardized dataset of Asian has not been reported before. Obviously, there are differences in the incidence of skin diseases, disease characteristics, and the background of skin color among different races.

2) Most of the images currently available in the dataset are watermarked, which may cause interference in the identification and analysis of skin lesions.

Table 1. Comparisons of clinical skin disease datasets.

Dataset	Classes	Amount	Region	Watermarking?	Available?
Dermofit	10	1,300	The UK	No	No
ISIC	16	23,906	Europe	No	Yes
Derm101	525	22,979	The US	No	Yes
Dermnet	626	18,974	Europe	Yes	Yes
DermIS	735	7,172	Germany	Yes	Yes
AtlasDerm	534	9,503	Brazil	Yes	Yes
Danderm	91	1,110	Denmark	Yes	Yes
Ours	541	107,565	China	No	Yes

In this paper, we establish a large-scale, Asian-dominated clinical image dataset of skin diseases, and carries out researches on it. The statistical data of XiangyaDerm is also presented in Table 1 for comparison with other public datasets. Our dataset will be publicly released for research purposes to the internet soon after, and the future update information could be found in this URL, http://airl.csu.edu.cn.

Fig. 1. Some sample images in XiangyaDerm. Each line from top to bottom are clinical images of basal cell carcinoma (BCC), pigmented nevus (PN), eczema (ECZ), lupus erythematosus (LE), lichen planus (LP), pemphigoid (PD), pemphigus (PS), psoriasis (PSO), squamous cell carcinoma (SCC), and seborrheic keratosis (SK).

3 Dataset

3.1 Data acquisition and cleaning

The collection of XiangyaDerm was approved by the Ethics Committee of Xiangya Hospital of Central South University, and informed consent was obtained from all participants. All clinical images were taken by dermatologists from Xiangya Hospital under standard illumination using four different cameras: SONY DSC-HX50 (350dpi), CANON IXUS 50 (180dpi), NIKON D40 (300dpi), NIKON COOLPIX L340 (300dpi), corresponding resolution of 3,888×5,184, 1,944×2,592, 2,000×3,008, 3,864×5,152. Finally, a total dataset of 47,075 images was obtained, covering 541 skin diseases, accounted for almost 99% of the incidence of skin diseases. The diagnostic labels for each image are validated by the gold standard of pathology and are supported by the patient's full medical history. We show some sample images in Figure 1. We can see that these images are with high image quality, simple background, and focus mainly on the typical skin lesions. For example, we chose the images of pemphigus with bullae instead of papules and plaques.

The data cleaning process is also accomplished by dermatologists from Xiangya Hospital. Five categories of images are removed in this process to obtain a clean dataset: Case 1: Images with low-quality due to improper shooting. Case 2: Images incorrectly labeled which confirmed to be inconsistent with the patient's medical history. Case 3: Skin lesion areas are covered by obvious local treatment or any other colored residues, which may have serious adverse effects on the training process. Case 4: Images contains obvious information about human body parts, such as nose, eyes, hair and so on, which can also interfere with the subsequent recognition. Case 5: Excessive exudate, which leads to the loss of the surface appearance and texture of the disease.

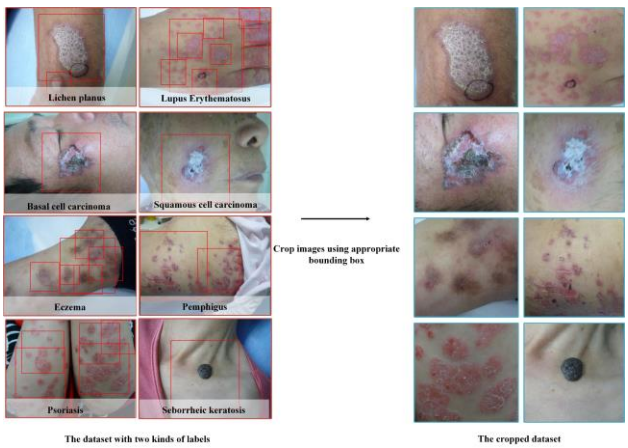

Fig. 2. The data annotation and cropping process for XiangyaDerm.

3.2 Data annotation and cropping

The annotation process is completed by 20 professional dermatologists, with more than 5 years of clinical work experience. These doctors were divided into two separate groups, each labeled half of the images and then cross-checked the other half. The task of annotation is to use labelImg, an open source image annotation software, to mark the typical lesion areas on the picture, that is, to represent the lesion area with a bounding box.

The cropping process uses the coordinates of the bounding box on the image to save that part of the image. As shown in Figure 2, after the cropping operation, not only the complex background of the skin image is removed, but also the amount of dataset is increased, since a picture may have several typical isolated skin lesions.

Eventually, the data volume of our dataset increased from 44,108 to 107,565, covering 541 categories of skin diseases, and the images were more concentrated on skin lesions. The largest amount of data in our dataset is psoriasis, 67,066 images, accounting for 62% of the total dataset. This is mainly because the dermatology department of Xiangya Hospital is a special outpatient department of psoriasis, and there are many patients with psoriasis every year. In addition to psoriasis, the data distribution of the remaining 540 skin diseases is shown in Figure 3, with the horizontal and vertical axes representing the disease and its corresponding data volume.

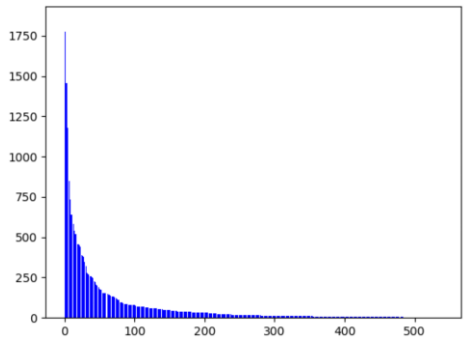

Fig. 3. The distribution of the final proposed dataset except psoriasis.

4 Experiment

4.1 80-classification experiment on XiangyaDerm

In order to evaluate the performances of different CNNs on this dataset and prove the usefulness of it, we select 4 mainstream CNN architectures, including InceptionV3[14], InceptionResNetV2[15], DenseNet121 [16] and Xception [17] to classify 80 common skin diseases. We select 80 kinds of skin diseases in our dataset whose amount of data is more than 100, and remove the parts whose amount of data is more than 1,000 in order to balance the chose 80 skin disease. The specific number of these

80 diseases in this experiment can be seen in the submitted supplementary files, which name is “appendix.pdf”.

In this experiment, our dataset is randomly divided into training set and test set in a ratio of 3:1. The whole training process was completed on 3 graphic cards of NVIDIA TITAN Xp. The image input size for InceptionV3, InceptionResNetV2 and Xception are both 299×299×3 and for Densenet121 is 224×224×3. We kept the rest of the experimental conditions consistent, for example, setting the same pretrained weights on ImageNet dataset, max training epochs 5000, basic learning rates 0.001, batch size 25, optimizer Adam, and the loss function categorical cross entropy. By organizing 4-fold cross validation experiments, we summarize the average values of the experimental results as shown in Table 2 and Figure 4.

Table 2. Recognition rate of 80-classification experiment.

Method	InceptionV3	DenseNet121	Xception	InceptionResNetV2
Top1 ACC	0.470	0.494	0.523	0.588
Top3 ACC	0.671	0.696	0.707	0.764

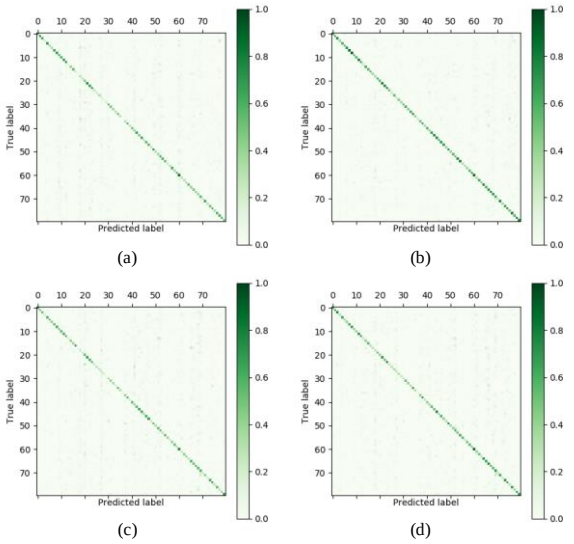

Fig. 4. Confusion matrices of 80-classification experiments: (a), (b), (c) and (d) represent test results for InceptionV3, DenseNet121, Xception, and InceptionResNetV2, respectively.

From the confusion matrix shown in Figure 4, we can see these 4 CNNs all have good performances. We can see that the dark green area of the confusion matrix is mainly distributed on the diagonal line, while the color of other areas is relatively light. It shows that most of the 80 skin diseases can be effectively distinguished by the adopted CNNs, except for a few indistinguishable diseases. However, it does not mean that

there are problems with our dataset or CNNs, which just shows that the diagnosis of these skin diseases based only on clinical images still faces challenges.

As we can see from Table 2, we can draw a preliminary conclusion that Inception-ResNetV2 could get better performance over other 3 networks in the 80-classification experiment on XiangyaDerm. Note that our goal is not to find the best network for recognition, but to verify the usefulness of our proposed datasets and give a baseline performance on it.

4.2 6-classification cross-test of Derm101 and XiangyaDerm

As mentioned earlier in this paper, we have been successfully approved by the Derm101 team to use their images for research purposes. The objective of this experiment is to obtain the cross-test performance of datasets between different races. We chose 6 common diseases witch high incidence and both occurred in Derm101 and XiangyaDerm whose amount is more than 100, including Basal Cell Carcinoma, Epidermoid Cyst, Psoriasis, Rosacea, Seborrheic Keratosis and Stasis Dermatitis. To balance each category in the datasets we took 100 images from each category and formed two sub-databases, namely Derm101-6 and Xiangya-6. Another dataset is Mix-6, which is a dataset composed of Derm101-6 and Xiangya-6.

As for the experimental settings, we used InceptionResNetV2, which is the best performing CNN in 80-classification experiment, to do cross-test on Derm101-6, Xiangya-6, and Mix-6. The loss function, batch size and other parameter settings are kept the same as the previous experiment. The cross-test here means that the model trained on one dataset is tested on the other two datasets. The test results obtained in this experiment are shown in Table 3 and Figure 5.

Table 3. Recognition rate of 6-classification experiment.

	Test on Derm101-6	Test on Xiangya-6
Train on Derm101-6	0.800	0.193
Train on Xiangya-6	0.213	0.720
Train on Mix-6	0.671	0.621

As we can see from (a) and (d) of Figure 5, the dark squares are concentrated on the diagonal lines which shows that the classification of each disease is also good, while the accuracy is much worse when we exchange the test sets. From Table 3, from the comparison between (a) (d) and (b) (c), we can see that the model trained and test on the same dataset has a better performance than the cross-test ones. As we can see from (a) and (d) of Figure 5. Comparing (e) (a) (c) and (f) (b) (d), we can easily find that the model trained on the mixed training dataset reached a better performance than training and test on a totally different dataset but worse than training and test on a same dataset.

From this, we can see that the cross-test performance of classification models between different races is not good. Through communication with professional dermatologists, we understand that there are differences in the incidence of skin diseases,

disease manifestations, and skin color among different races, which can lead to the failure of classification models. Therefore, to build a skin disease CAD system with high performance and stability, we recommend to establish a specific dataset of skin diseases for different regions and races.

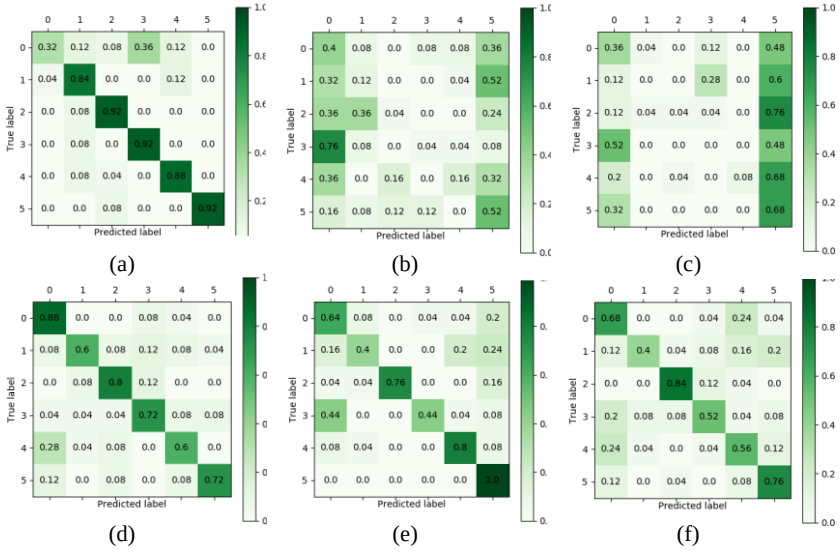

Fig. 5. Confusion matrices of 6-classification experiment: (a) Derm101-6 training Derm101-6 test, (b) Derm101-6 training Xiangya-6 test, (c) Xiangya-6 training Derm101-6 test, (d) Xiangya-6 training Xiangya-6 test, (e) Mix-6 training Derm101-6 test, (f) Mix-6 training Xiangya-6 test.

5 Conclusion

In this paper, we propose a clinical image dataset for Asian race's skin disease diagnosis system. It contains 107,565 images, ranging from 541 categories. Each image is confirmed by a disease label and is marked by a specialist with bounding boxes. Our experiments demonstrate the classification performances of the current state-of-the-art CNN architectures and demonstrate its usability as a benchmark dataset for the diagnosis of skin diseases. Moreover, we have also proved that it is necessary to construct a specific dataset of skin diseases for different regions and races through the cross-test experiments. The XiangyaDerm proposed in this paper can effectively promote the research and application of Asian skin disease diagnosis, and is also a useful supplement to global skin data.

References

1. Hay R J, Johns N E, Williams H C, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions[J]. Journal of Investigative Dermatology, 2014, 134(6): 1527-1534.
2. Gonzalez-Castro V, Debayle J, Wazaefi Y, et al. Automatic classification of skin lesions using color mathematical morphology-based texture descriptors[C]//Twelfth International Conference on Quality Control by Artificial Vision 2015. International Society for Optics and Photonics, 2015, 9534: 953409.
3. Badano A, Revie C, Casertano A, et al. Consistency and standardization of color in medical imaging: a consensus report[J]. Journal of digital imaging, 2015, 28(1): 41-52.
4. Sun X, Yang J, Sun M, et al. A benchmark for automatic visual classification of clinical skin disease images[C]//European Conference on Computer Vision. Springer, Cham, 2016: 206-222.
5. Yang J, Sun X, Jie L, et al. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018, 11.
6. Liao H. A deep learning approach to universal skin disease classification[R]. University of Rochester Department of Computer Science, CSC, 2016.
7. Liao H, Li Y, Luo J. Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks[C]//Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016: 355-360.
8. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee, 2009: 248-255.
9. Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.
10. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.
11. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
12. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
13. Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115.
14. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.
15. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//AAAI. 2017, 4: 12.
16. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//CVPR. 2017, 1(2): 3.
17. Chollet F. Xception: Deep learning with depthwise separable convolutions[J]. arXiv preprint, 2017: 1610.02357.