Journal of Medical Systems (2019) 43: 283
https://doi.org/10.1007/s10916-019-1414-2

PATIENT FACING SYSTEMS

The Application of Deep Learning in the Risk Grading of Skin Tumors for Patients Using Clinical Images

Xin-yu Zhao[1] · Xian Wu[2] · Fang-fang Li[3,4,5] · Yi Li[1] · Wei-hong Huang[6] · Kai Huang[3,4,5] · Xiao-yu He[1] · Wei Fan[2] · Zhe Wu[1] · Ming-liang Chen[3,4,5] · Jie Li[3,4,5] · Zhong-ling Luo[3,4,5] · Juan Su[3,4,5] · Bin Xie[1] · Shuang Zhao[3,4,5]

Received: 10 May 2019 / Accepted: 3 July 2019 / Published online: 13 July 2019

Abstract

According to diagnostic criteria, skin tumors can be divided into three categories: benign, low degree and high degree malignancy. For high degree malignant skin tumors, if not detected in time, they can do serious harm to patients' health. However, in clinical practice, identifying malignant degree requires biopsy and pathological examination which is time costly. Furthermore, in many areas, due to the severe shortage of dermatologists, it's inconvenient for patients to go to hospital for examination. Therefore, an easy to access screening method of malignant skin tumors is needed urgently. Firstly, we spend 5 years to build a dataset which includes 4,500 images of 10 kinds of skin tumors. All instances are verified pathologically thus trustworthy; Secondly, we label each instance to be either low-risk, high-risk or dangerous in which Junctional nevus, Intradermal nevus, Dermatofibroma, Lipoma and Seborrheic keratosis are low-risk, Basal cell carcinoma, Bowen's disease and Actinic keratosis are high-risk, Squamous cell carcinoma and Malignant melanoma are dangerous; Thirdly, we apply the Xception architecture to build the risk degree classifier. The area under the curve (AUC) for three risk degrees reach 0.959, 0.919 and 0.947 respectively. To further evaluate the validity of the proposed risk degree classifier, we conduct a competition with 20 professional dermatologists. The results showed the proposed classifier outperforms dermatologists. Our system is helpful to patients in preliminary screening. It can identify the patients who are at risk and alert them to go to hospital for further examination.

Keywords Convolutional neural network · Skin tumor · Risk grade · Preliminary screening

Introduction

Skin tumor is a common disease and most of them are benign. However, a small fraction of skin tumors could turn to malignant skin cancers which are fatal to patients [1]. Take melanoma for example, although it only makes up 1% of all skin cancers, melanoma processes an extremely high fatality rate [2] and becomes a major public health concern [3–7]. Besides melanoma, there're also other high-risk skin cancers, like squamous cell carcinoma which is also a common malignant neoplasm. The incidence rate of skin cancers is increasing worldwide [1], only in US, more than 5 million people are diagnosed to be skin cancer each year. Due to the severe impact and high incidence rate of skin cancer, evaluating the malignancy degree of skin tumors become an important task.

In clinical setting, the golden diagnosis of benign and malignant skin tumors requires biopsy and pathological examination, which are time-consuming and costly for patients. Moreover, in China, there is a severe shortage of dermatologists, the dermatologist and patient ratio is as large as 1:60000.

Xin-yu Zhao, Xian Wu and Fang-fang Li contributed equally to this work.

This article is part of the Topical Collection on *Patient Facing Systems*

✉ Bin Xie
xiebin@csu.edu.cn

✉ Shuang Zhao
shuangxy@csu.edu.cn

[1] School of Automation, Central South University, Changsha, China

[2] Tencent Medical AI Lab, Beijing, China

[3] Department of Dermatology, Xiangya Hospital Central South University, Changsha, China

[4] Hunan Key Laboratory of Skin Cancer and Psoriasis, Changsha, China

[5] Hunan Engineering Research Center of Skin Health and Disease, Changsha, China

[6] Mobile Health Ministry of Education - China Mobile Joint Laboratory, Xiangya Hospital Central South University, Changsha, China

In the rural area of China, there are few dermatologists. Therefore, it's very inconvenient for patients to go to hospital, as they may need to go to superior hospital in big city and wait a long queue for examination. In this case, if there is an easy-to-access preliminary screening method, we can identify the patients who are at risk and alert them to go to hospital for further examination. For example, if the skin lesions are recognized as benign, there is no need for patients to worry, just keep monitoring the condition. If the skin lesions are recognized as low-malignant, it may be necessary for patients to go to nearby hospital for confirmation. If the skin lesions are recognized as high-malignant, patients need to go to superior hospital urgently. In this case, patients can avoid delaying treatment because of no awareness of malignant melanoma, or worrying too much about their lives because of a pigmented plaque, while it only is pigmented nevus. Therefore, design an automatic grading system for skin tumors will bring great convenience to patients which lack of adequate medical knowledge.

To help patients to conduct preliminary screening, in this paper, we explore the deep neural network to automatically grade the malignancy of skin tumors. Recently, the rapid development in deep learning have made great achievements in many computer vision tasks. Deep learning is a class of machine learning algorithms which use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. The convolutional neural network (CNN) is one of the deep learning models and have achieved great empirical successes in field of computer aided diagnosis, such as diabetic retinopathy [8], genetic disorders [9] and radiology [10]. Dermatology is another case demonstrating promising results by using CNNs. Liao H et al. trained multi-class CNN classifiers for universal skin diseases, achieving high accuracies [11–13]. Sun X et al. have attempted to classified 128 skin diseases by using CNNs [14]. Moreover, skin tumors have been focused on particularly by many researches in this filed because of its fatalness. For instance, [15–17] were based on dermoscopic images and [18, 19] were based on clinical images. As a landmark publication, Esteva et al. used 129450 images to train a binary (benign/malignant) classifier and the performance can reach the same level as that of board-certified dermatologists [18]. Haenssle et al. demonstrated that the CNN's diagnostic performance was superior to most but not all dermatologists, in conjunction with results from the reader study level-I and -II (level-I: dermoscopy only; level-II: dermoscopy plus clinical information and images) [20]. Walker et al. proposed a novel two-stage bedside skin cancer diagnosis system and achieved a high accuracy in a prospective study [21]. However, all studies are about the classification of specific species of skin tumors but automatic grading of malignant degree for patients to conduct initial screening by themselves has not been well explored.

In this paper, we propose a deep learning model to evaluate the risk level of skin tumors. Firstly, since the quality of deep learning model heavily relies on the quality of training data, we spend 5 years to collect data from in Xiangya hospital which is top hospital in China. This data set includes 10 kinds of skin tumors and each instance is verified pathologically, thus the label is trustworthy; Secondly, we design a classifier which could grade malignant degree into three levels: low-risk, high-risk and dangerous. Our model can reach the accuracy of 82.7%, and the ROC for these risk degrees reach 0.959 for low-risk, 0.919 for high-risk and 0.947 and dangerous respectively. To further verify the performance of the proposed classifiers, we conduct a competition with 20 professional dermatologists in which the proposed risk degree classifier outperforms professional dermatologists. Based on our classification results, for low-risk, we recommend patients keep monitoring condition; For high-risk, we recommend patients go to nearby hospital; For dangerous, we recommend patients take pathological examination in superior hospital immediately.

Material and methods

The dataset was considered for use in this study, named as XiangyaDerm, which was collected from Xiangya Hospital by using professional digital camera (Canon, Resolution: 350dpi). A total of 150,223 clinical images taken from 2014 through 2018 were contained in this dataset. This dataset involved 543 skin diseases, from which we selected 10 types of skin tumors for studying grading of malignant degree, considering data volume and data balance. The sample of 10 types of skin tumors is shown in Fig. 1. Each image has a corresponding medical history and pathological diagnosis.

Data filter

We conduct a data pre-processing over collected images before training a classifier. In particular, the images of skin lesions coated by colored potion may mislead CNNs to learn the incorrect color characteristic of a disorder. The lesions of some special parts like finger terminal have not enough area to represent more feature, and most images of them have more background than skin area. Hence, we remove them from dataset. Other types of images like skin lesions covered by hair and excessive exudate on the skin surface was also considered to be abandoned.

Annotation

Annotation is conducted by professional dermatologists. The ground truth labels need to be verified by pathological examination and confirmed by the information of medical record and doctor's experience accumulate over a long period of time. Two types of annotation containing bounding box and

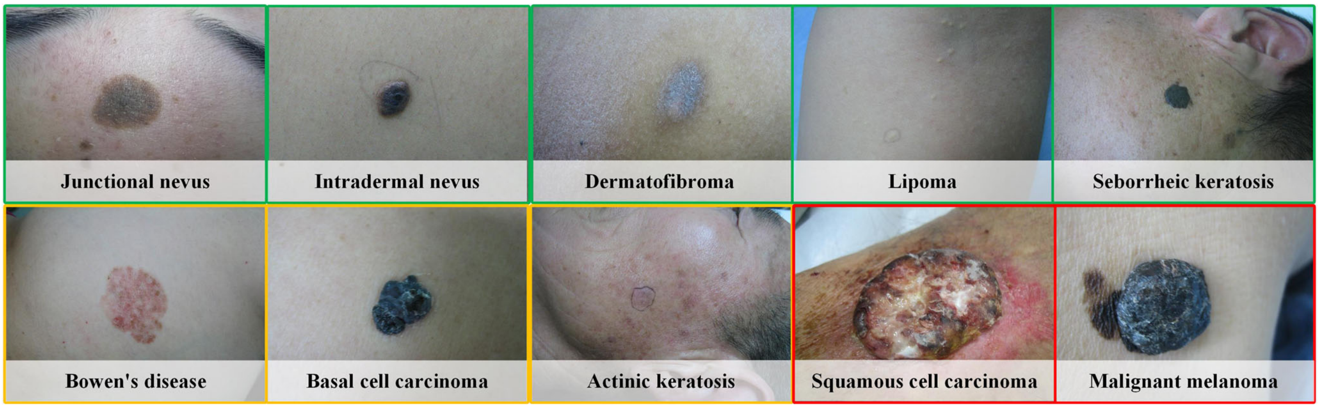


Fig. 1 The sample of 10 types of skin tumors in our dataset. The images with green boxes in the first row are from low-risk diseases; the images with yellow boxes in the second row are from high-risk diseases; the images with red boxes in the second row are considered from diseases

the name of its class were provided by using labeling tool. Hence, a standardized dataset could be accessed for detection and classification tasks.

Method

The core algorithm of our system is deep learning method. The Google team proposed GoogLeNet in 2014 and then proposed various modifications, further producing networks, such as Inception v3, InceptionResNet V2, and Xception, with a lower error rate. Xception adopts depth-wise separable convolution to replace the convolution action of the original Inception v3 and improves the network performance. We fine-tuned the Xception pretrained on ImageNet to complete the task of soring degree. Software tools for deep learning were available widely, which we chose was Tensorflow (https://www.tensorflow.org; Google Brain Team, Mountain View, CA).

The training of deep learning method can be understood as the process of super-high dimension parameter fitting. Cross entropy as the loss function of the model was used to depict the distance between the real label of the picture and its predicted label. The formula was expressed as follows:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n log\hat{y}_n + (1-y_n)\log\left(1-\hat{y}_n\right)\right]$$

Where N is the number of sample in a batch, y_n is the true label, $\hat{y}_n$ is predicted label. In order to achieve a smaller loss function as soon as possible, a reasonable weight updating strategy needs to be set up. We used RMSprop as the

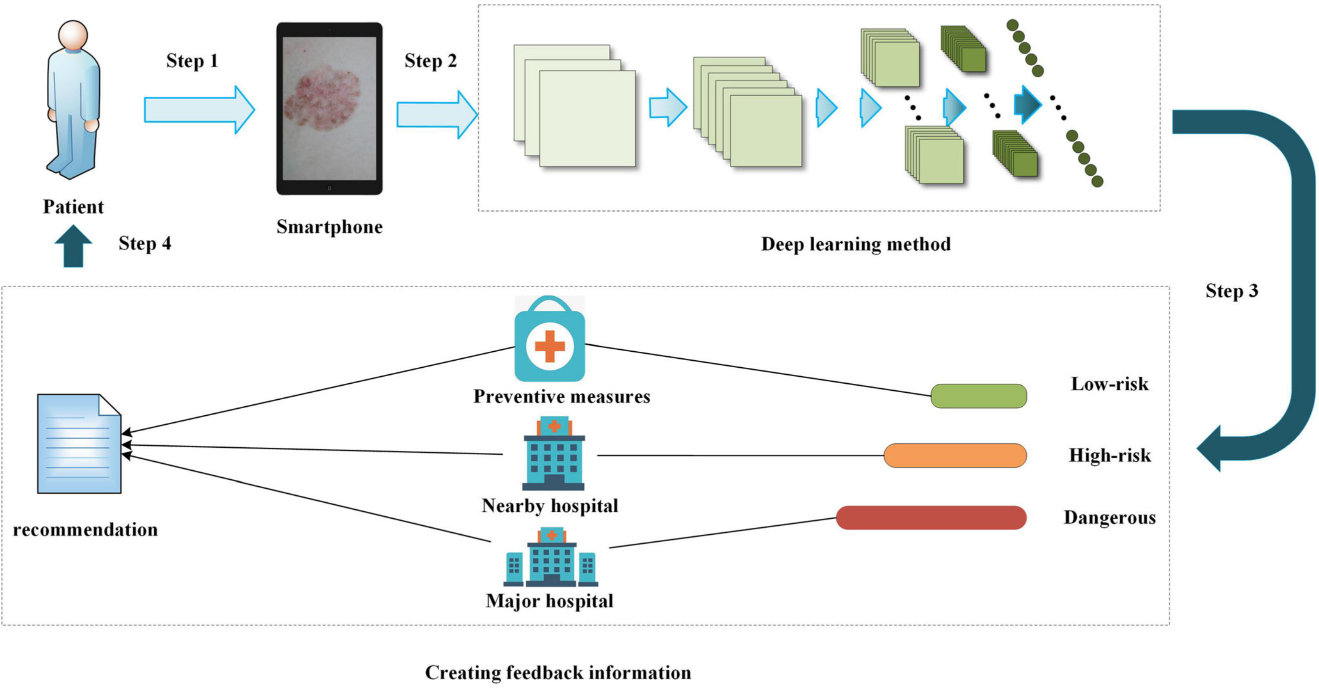


Fig. 2 Procedure of risk grading system for skin tumor based on deep learning

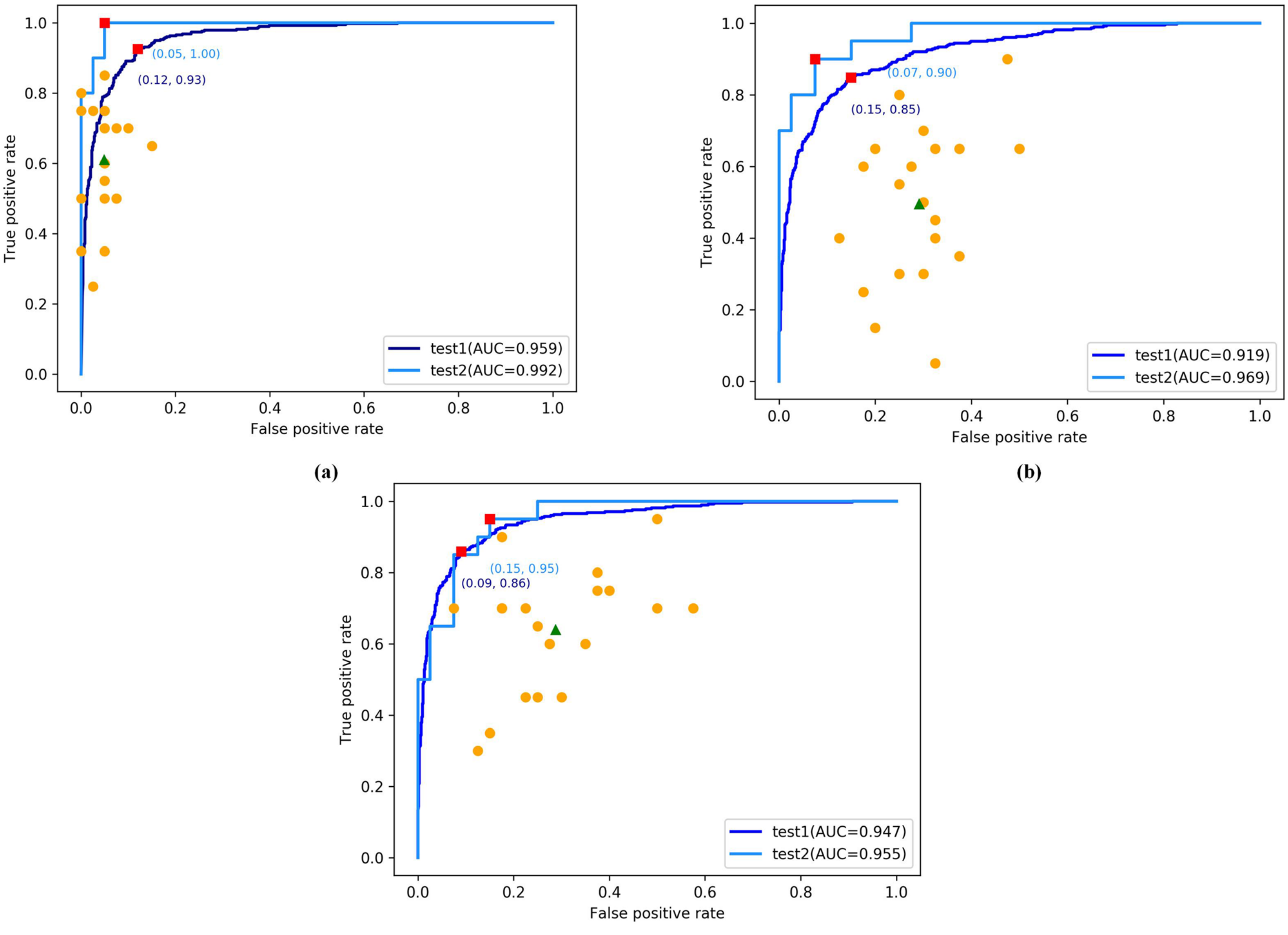
optimizer, decay rate was set as 0.9, learning rate was set as 0.001. And epoch was set as 100 for better training effect.

The test set was chosen from 4500 images in accordance with a ratio of 1/4 of the entire data set. The rest were used for training. We input images into networks in batches and the batch size was set to be 25 considering the computing power of graphics card and the fitting efficiency of the algorithm. To avoid similar pictures in same batch, we shuffled the order of the data in the import process.

The procedure of our system can be summarized into four steps as shown in Fig. 2. In step one, patients are asked to obtain clinical images using smartphones, which are accessible device for most people. In step two, the captured images were submitted by a purpose-built application to the classifier trained by deep learning method, and core algorithm often operate in the cloud. In step three, three degrees of malignancy are output. According to textbooks and medical guidelines, we define Junctional nevus, Intradermal nevus, Dermatofibroma, Lipoma and Seborrheic keratosis as benign, Basal cell carcinoma, Bowen's disease and Actinic keratosis as low-malignant, Squamous cell carcinoma and Malignant melanoma as high-malignant. When predicted skin tumor is benign, we regard it as low-risk to health. There is no need to worry too much about condition for patients, and they only need to keep monitoring the condition and take some preventive measures such as reduce sun exposure. When predicted skin tumor is low-malignant, we regard it as high-risk to health. Patients are suggested to visit a nearby hospital; If predicted skin tumor is high-malignant, it will dangerous for patients to do nothing. They are urgent to do pathological examination in major hospital. In the last step, all predicted results and corresponding recommendation will be sent to the uploader.

Results

In this section, we conduct two tests to evaluate performance. Firstly, our classifier was test on 1125 clinical images. Then,



(a) **(b)** **(c)**

Fig. 3 Receiver operating characteristic (ROC) curves for malignant degree grading. Test1(dark blue curve) was tested with 1125 images using Xception; test2 using Xception (dodger blue curve) and 20 dermatologists (orange dot 20 dermatologists; green triangle average value of 20 dermatologists) were tested with 60 chosen images from the Xiangya dataset. (a)Benign. (b)Low degree. (c)High degree

sensitivity which represents the proportion of correct prediction in positive samples and specificity which represents the proportion of correct prediction in negative samples were calculated. And both of them could reflect the condition of missed diagnosis and misdiagnosis. In our prediction system, the output class of networks determined by whether exceeding the set threshold. By adjusting the threshold, we could obtain a series of sets consisting of sensitivity and specificity, and receiver operating characteristic curves (ROC) were shown in Fig. 3. The area under the curve (AUC) for the grade of benign, low degree and high degree were 0.959, 0.919 and 0.947. Youden index is defined for all points of ROC curve (sensitivity + specificity −1). We used maximum value of the index for selecting the optimum threshold and results. From the Fig. 3, we could know that sensitivity (1 – the probability of missed diagnosis) and specificity (1 – the probability of misdiagnosis) were 0.93, 0.88 for benign, low degree of those were 0.85, 0.85, and high degree of those were 0.86, 0.91. It indicated that our Artificial intelligence (AI) system has a very small probability of misdiagnosis and missed diagnosis to three malignant degrees. In this test, the overall accuracy could achieve 0.827, which also indicated our classifier have good performance.

For practical purposes, 60 test images were chosen from Xiangya dataset to compare the performances of the AI system with 20 dermatologists. Each grade contained 20 pictures. The false positive rate and true positive rate of every dermatologist were plotted and shown in Fig. 3. For 20 dermatologists, Average value of false positive rate and true positive rate were (0.049, 0.61), (0.29, 0.495) and (0.29, 0.64), which were worse than deep learning method. We found that most of these points are below the curve, which indicated our AI system could outperform the dermatologists in the Malignant degree grading. Moreover, Fig. 4 shown the confusion matrix of our method in comparison to the 20 tested dermatologists. Element (i, j) of each confusion matrix represents the empirical probability of predicting class j given that the ground truth was class i. We could get the result from the figure that low degree and high degree were easier to be confused than benign by both AI system and dermatologists. However, AI system have better overall performance.

Discussion

The low accuracy of human dermatologists can be explained. In clinical practice, there is no strict criteria for the grades of skin tumors malignant degree, and there is no clear boundary for the different degrees. Hence, some doctors will have different grading results for skin diseases being on the boundary and deviate from our preset boundaries, so the final result is that benign diseases are easy to be confused with low-malignant skin diseases, low-malignant diseases are easy to be confused with high-malignant diseases, while high-malignant diseases are difficult to confuse with benign diseases. However, our AI system is mainly used in the initial screening for patients, we don't need to accurately diagnose individual diseases. As long as it has a good grading result under a given boundary, the system will help most patients and doctors.

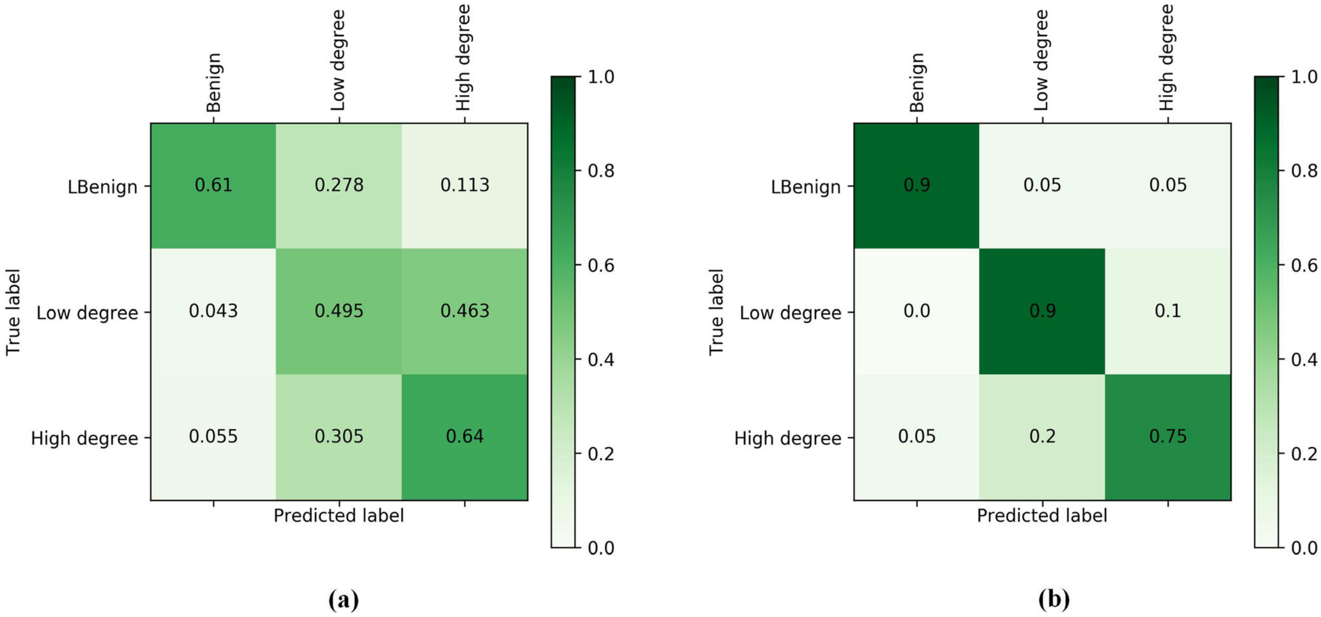


Fig. 4 Confusion matrices on test set which contains 60 chosen images from the Xiangya dataset. (a) average result of 20 dermatologists. (b) Deep learning method

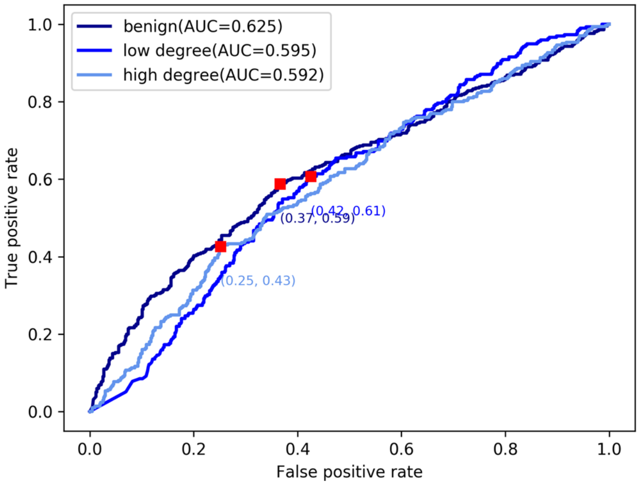


Fig. 5 Receiver operating characteristic (ROC) curves for malignant degree grading tested on dataset consisting of other racial patients

The training set in this paper does not cover all types of skin tumors. In order to verify whether it has the function of preliminary malignant screening when encountering diseases not contained in training set. We selected 540 pictures of hemangiomas and syringomas belonging to benign dermatosis as the new test set. The experimental results show that the accuracy of our AI system can reach 0.72. We supposed that our algorithm can learn the uniform depth features of specific malignant degrees. Even if we encounter other diseases not included in the training set, we can grade the malignant degree with a higher accuracy. The corresponding APP have been working on. We hope that in the future, patients in areas with insufficient medical resources can utilize this APP to determine the malignant degree initially, and according to the results, they can consider whether there is an urgent need to do biopsy and pathological examination.

However, this AI system still has obvious shortcomings. We tried to test our system with datasets from other races. A total of 1015 images were selected from Atlas Derm dataset, DermIS dataset and Derm101 dataset and divided into three malignant levels to build a new test set. The ROC curve is shown in Fig. 5. Compared with the previous results, the performance on this dataset have obviously decreased. We supposed that different races have contributed to this situation. In fact, similar phenomena have occurred in Han S S et al. Therefore, for better generalization performance, it will be important to increase the number of available clinical images of patients of different ages and ethnicities in future study.

Conclusions

In this paper, we built a new dataset accumulated from Xiangya Hospital Central South University for malignant degree grading. All images were annotated by professional dermatologists and verified pathological. We have divided the risk grades into three levels through the degree of malignancy: low-risk, high-risk and dangerous. The experimental results and human-versus-machine competitions prove that our system is capable of automatically grading a given clinical image of skin tumors. This is the first time that deep learning method have been attempted to be used in automatic risk grading for skin tumors. These results may have implications in the computer aided systems for skin diseases and our system are helpful to patients in preliminary screening before diagnosis by themselves.

Acknowledgements First, second and third authors contributed equally to this paper. Authors are grateful to all the doctors and nurses in Department of Dermatology, Xiangya Hospital Central South University.

Funding This research was funded by National Key R&D Program of China (2018YFC0117000), Specialized Basic Work of Science and Technology (2015FY111100) and Hunan Provincial Science and Technology Department (2018SK2092).

Compliance with Ethical Standards

All the authors of this article are aware of the content.

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Apalla, Z., Lallas, A., Sotiriou, E., Lazaridou, E., and Ioannides, D., Epidemiological trends in skin cancer. Dermatol Pract Concept. 7(2):1, 2017.
2. Wernli, K. J., Henrikson, N. B., Morrison, C. C., Nguyen, M., Pocobelli, G., and Blasi, P. R., Screening for skin cancer in adults: Updated evidence report and systematic review for the US preventive services task force. JAMA. 316(4):436–447, 2016.
3. Zhao, S., Wu, L., Kuang, Y., Su, J., Luo, Z., Wang, Y., Li, J., Zhang, J., Chen, W., Li, F., and He, Y., Downregulation of CD147 induces malignant melanoma cell apoptosis via the regulation of IGFBP2 expression. Int J Oncol. 53(6):2397–2408, 2018.
4. de Polo, A., Luo, Z., Gerarduzzi, C., Chen, X., Little, J. B., and Yuan, Z. M., AXL receptor signalling suppresses p53 in melanoma through stabilization of the MDMX–MDM2 complex. J Mol Cell Biol. 9(2):154–165, 2017.
5. Liu, X. S., Genet, M. D., Haines, J. E., Mehanna, E. K., Wu, S., Chen, H. I., Chen, Y., Qureshi, A. A., Han, J., Chen, X., and Fisher, D. E., ZBTB7A suppresses melanoma metastasis by transcriptionally repressing MCAM. Mol Cancer Res. 13(8):1206–1217, 2015.
6. Zeng, W., Su, J., Wu, L., Yang, D., Long, T., Li, D., Kuang, Y., Li, J., Qi, M., Zhang, J., and Chen, X., CD147 promotes melanoma progression through hypoxia-induced MMP2 activation. Curr Mol Med. 14(1):163–173, 2014.
7. Luo, Z., Zeng, W., Tang, W., Long, T., Zhang, J., Xie, X., Kuang, Y., Chen, M., Su, J., and Chen, X., CD147 interacts with NDUFS6 in regulating mitochondrial complex I activity and the

mitochondrial apoptotic pathway in human malignant melanoma cells. Curr Mol Med. 14(10):1252–1264, 2014.

8. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., and Kim, R., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA – J Am Med Assoc. 316(22):2402–2410, 2016.
9. Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P. M., Kamphausen, S. B., Zenker, M., and Bird, L. M., Identifying facial phenotypes of genetic disorders using deep learning. Nature medicine. 25(1):60, 2019.
10. Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., and Karssemeijer, N., Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 35:303–312, 2017.
11. Liao, H., A deep learning approach to universal skin disease classification. University of Rochester Department of Computer Science, CSC, 2016.
12. Liao, H., Li, Y., Luo, J., Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks. In 2016 23rd International Conference on Pattern Recognition (ICPR) 2016 Dec 4 (pp. 355-360). IEEE.
13. Haofu, L, Luo, J., A deep multi-task learning approach to skin lesion classification. InWorkshops at the Thirty-First AAAI Conference on Artificial Intelligence 2017 Mar 21.
14. Sun, X., Yang, J., Sun, M., Wang, K., A benchmark for automatic visual classification of clinical skin disease images. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 206-222). Springer, Cham.
15. Codella, N. C., Nguyen, Q. B., Pankanti, S., Gutman, D. A., Helba, B., Halpern, A. C., and Smith, J. R., Deep learning ensembles for melanoma recognition in dermoscopy images. IBM Journal of Research and Development. 61(4/5):5–1, 2017.
16. Yu, L., Chen, H., Dou, Q., Qin, J., and Heng, P. A., Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging. 36(4):994–1004, 2017.
17. Zhang, J., Xie, Y., Wu, Q., Xia, Y., Skin lesion classification in dermoscopy images using synergic deep learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2018 Sep 16 (pp. 12-20). Springer, Cham.
18. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S., Dermatologist-level classification of skin cancer with deep neural networks. Nature. 542(7639):115, 2017.
19. Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., and Chang, S. E., Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol. 138(7):1529–1538, 2018.
20. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B., Thomas, L., Enk, A., and Uhlmann, L., Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 29(8):1836–1842, 2018.
21. Walker, B. N., Rehg, J. M., Kalra, A., Winters, R. M., Drews, P., Dascalu, J., David, E. O., Dascalu, A., Dermoscopy diagnosis of cancerous lesions utilizing dual deep learning algorithms via visual and audio (sonification) outputs: Laboratory and prospective observational studies. EBioMedicine, 2019.