Assessing the effectiveness of artificial intelligence methods for melanoma: A retrospective review

Xiaoyu Cui, PhD,[a] Ran Wei, BA,[a] Lixin Gong, MS,[a] Ruiqun Qi, MD,[b] Zeyin Zhao, MS,[a] Hongduo Chen, MD,[b] Kaixin Song, BA,[a] Amer A. A. Abdulrahman, MD,[b] Yining Wang, BM,[b] John Z. S. Chen, MD,[c] Shuo Chen, PhD,[a] Yue Zhao, PhD,[a] and Xinghua Gao, MD[b]
Shenyang, China, and Tucson, Arizona

Background: Artificial intelligence methods for the classification of melanoma have been studied extensively. However, few studies compare these methods under the same standards.

Objective: To seek the best artificial intelligence method for diagnosis of melanoma.

Methods: The contrast test used 2200 dermoscopic images. Image segmentations, feature extractions, and classifications were performed in sequence for evaluation of traditional machine learning algorithms. The recent popular convolutional neural network frameworks were used for transfer learning training classification.

Results: The region growing algorithm has the best segmentation performance, with an intersection over union of 70.06% and a false-positive rate of 17.67%. Classification performance was better with logistic regression, with a sensitivity of 76.36% and a specificity of 87.04%. The Inception V3 model (Google, Mountain View, CA) worked best in deep learning algorithms: the accuracy was 93.74%, the sensitivity was 94.36%, and the specificity was 85.64%.

Limitations: There was no division in the severity of melanoma samples used in this experiment. The data set was relatively small for deep learning.

Conclusion: The performance of traditional machine learning is satisfactory for the small data set of melanoma dermoscopic images, and the potential for deep learning in the future big data era is enormous. (J Am Acad Dermatol 2019;81:1176-80.)

Key words: artificial intelligence; classification; deep learning; melanoma diagnosis; segmentation; traditional machine learning.

Globally, melanoma has the highest mortality rate of all skin cancers.[1] The early diagnosis of melanoma can significantly reduce the mortality[2] and complications.[3] In recent years, traditional machine learning (ie, machine learning before deep neural networks) has achieved satisfactory performance in the field of medicine, which is the trend toward future development.[4] In addition,

From the Department of Biomedical Informatics, College of Medicine and Biological Information Engineering, Northeastern University, Shenyang[a]; NHC Key Laboratory of Immunodermatology (China Medical University), Ministry of Education Key Laboratory of Immunodermatology (China Medical University), Department of Dermatology The First Hospital of China Medical University, Shenyang[b]; and HealthySkin Dermatology, LLP, Tucson.[c]

Authors Cui and Wei contributed equally to this article.

Funding sources: Supported by the National Natural Science Foundation of China (71621061, 61605025, 81673070, 81872538), the 111 Project (B16009, D18011), and the Fundamental Research Funds for the Central Universities (N171904006, N171902001, N172410006-2).

Conflicts of interest: None disclosed.

Accepted for publication June 19, 2019.

Reprint requests: Ruiqun Qi, MD, No.155 Nanjing Bei St, Heping District, 110001 Shenyang, China. E-mail: xiaoqiliumin@163.com.

Published online June 27, 2019.

0190-9622/$36.00

https://doi.org/10.1016/j.jaad.2019.06.042

artificial intelligence technology, represented by neural networks, has also achieved remarkable results in the diagnosis of melanoma.

At present, for the diagnosis of melanoma, there is a lack of comparative analysis of deep learning and traditional machine learning based on the same data set, for the purpose of choosing the optimal classifier or segmentation algorithm using traditional machine learning and the optimal model using deep learning, and inferring their advantages and disadvantages.

The objective of this study was to seek the best artificial intelligence based on the same data set. This traditional machine learning part involves 4 steps: (1) image preprocessing, (2) image segmentation, (3) feature extraction, and (4) classification. Ultimately, the goal of this report is to demonstrate that the region growing algorithm has achieved good results in the segmentation of dermoscopic images. Support vector machine (SVM) and logistic regression (LR) are both suitable for the classification of melanoma dermoscopic images. In the deep learning part, the Inception V3 model (Google, Mountain View, CA) works best.

CAPSULE SUMMARY

- Some frequently used artificial intelligence algorithms were implemented and evaluated with the same data set for distinguishing melanoma.
- Region growing segmentation incorporated with logistic regression classifier can achieve a better performance in machine learning algorithms.
- Inception V3 (Google, Mountain View, CA) incorporated with transfer learning can achieve a better performance in deep learning algorithms.

METHODS

In this research, the traditional machine learning and deep learning algorithms were conducted in 2 parts, and 4 experiments were included. In the traditional machine learning experiment, we compared 4 image segmentation algorithms with the ground truths that were implemented by 3 dermatologists and then rechecked by a senior dermatologist. The best segmentation results were used for feature extraction, and then, 4 melanoma classifiers were compared. In the deep learning section, comparative experiments were conducted on 4 models that have become popular in recent years, and the optimal model for melanoma diagnosis was selected. Finally, the analysis and summary are given for the classifiers, the segmentation algorithms, and deep learning models.

The official dermoscopic image data set from the International Society for Digital Imaging of the Skin was used to test each method with a small data set and a large data set. There are 2200 images in the large data set, of which 564 are melanomas and 1636 are nonmelanomas. A subdata set comprised 606 images that were randomly selected, which included 295 melanomas and 311 nonmelanomas. The experiment environment was the Windows 7 system (Microsoft, Redmond, WA), the platform was MATLAB R2018b (MathWorks, Natick, MA) and TensorFlow 1.3 (Google), and the graphics processing unit was GTX1080Ti (NVIDIA, Santa Clara, CA).

In the machine learning part, the color images were converted to grayscale, and median filtering was used to reduce random noise. After that, the images were cropped to remove the black edges, and histogram equalization was performed to enhance the image contrast.[5]

After preprocessing, 4 fully automatic image segmentation algorithms—active contour model,[6] cluster,[7] region growing,[8] and Otsu[9]—were implemented, which were commonly used for segmenting melanoma. Dilating, corrosion, and holes filling were conducted as postprocessing. Finally, intersection over union and the false-positive rate were used to evaluate the segmentation results with the ground truths.

Color, morphologic, and texture features were extracted from the region of interest of the best performance segmentation algorithm. The mean, standard deviation, and skewness of 3 kinds of color space (RGB, HSV and Lab) were used as the color features[10]; morphologic features included the area, perimeter, roundness, and centroid, among others; the gray-level co-occurrence matrix, gray-level run length matrix, gray-level−gradient co-occurrence matrix, neighboring gray-level dependence matrix, grayscale histogram features, and Tamura features were calculated as the texture features.[11] In addition, the least absolute shrinkage and selection operator[12] method was used to select the most useful features from the primary data set. A score (Rad score) was calculated for each image via a linear combination of selected features that were weighted by their respective coefficients.

Four commonly used classifiers for computer-aided diagnosis of melanoma, including SVM,[13] classification and regression tree,[14] k-nearest neighbors,[15] and LR[16] were used for comparison. The

Abbreviations used:

LR: logistic regression
SVM: support vector machine
VGG: Visual Geometry Group network

accuracy, specificity, and sensitivity of these algorithms were calculated by the 10-fold cross-validation method.

In the deep learning part, because deep learning requires a large amount of image data, skin cancer images were fewer than needed. We thus used transfer learning to solve this problem. In transfer learning, the learned and trained model parameters are transferred to a new model to help the new model to train.[17] Considering that most of the data or tasks are relevant, we can learn the model parameters that we have learned through transfer learning. Sharing the new model in a way that speeds up and optimizes the learning efficiency of the model does not require learning from zero, as in most networks. The current commonly used AlexNet,[18] Visual Geometry Group network 16 (VGG16),[19] Visual Geometry Group network 19 (VGG19), Google Inception V3,[20] and other models are used for transfer learning training classification.

RESULTS

Segmentation algorithms comparison results

To evaluate the segmentation precision of these 4 algorithms, we compared their segmentation result with the region of interest labeled by dermatologists. As reported in Table I, for the small data sets that consist of 606 images, region growing achieves the best segmentation results, with a final average intersection over union of 70.06% and false-positive rate of 17.67%. Thus, the segmentation results of region growing were chosen as the criterion of feature extraction and classification.

Feature extraction and dimension reduction results

A total of 82 features were extracted from the marked region of interest of the region growing algorithm, which including 27 color features, 16 morphologic features, and 39 texture features. These 82 features were then reduced to 17 key features by the least absolute shrinkage and selection operator algorithm, which are listed in Table II.

Classifier comparison results

As summarized in Table III, the SVM algorithm and the LR algorithm can both achieve higher accuracy results (>81%). However, even though the specificity of SVM was up to 91%, sensitivity was only 71%. Accordingly, the sensitivity and specificity of LR were 76% and 87%, respectively. Thus, the LR algorithm was more suitable for the classification of melanoma dermoscopic images. In addition, the k-nearest neighbors algorithm can achieve a relative higher sensitivity compared with the other 3 methods.

Table I. Segmentation of melanoma by each segmentation algorithm

Segmentation algorithm	IOU, %	FPR, %
Active contour model	51.95	44.89
Cluster	33.32	45.43
Region growing	70.06	17.67
Threshold (Otsu threshold)	57.14	39.73

FPR, False-positive rate; *IOU*, intersection over union.

Table II. Feature extraction results by the least absolute shrinkage and selection operator algorithm

Feature type	Selected features
Color features	Skewness of blue channel (RGB color space), skewness of green-red channel (Lab color space), standard deviation of blue-yellow channel (Lab color space), standard deviation of hue channel (HSV color space), mean of saturation channel (HSV color space)
Morphologic features	Perimeter, solidity
Texture features	
GLRLM	Run length ratio, short run emphasis, long run emphasis, gray level distribution
GLCM	Energy standard deviation, contrast standard deviation, correlation standard deviation, contrast
GLGCM	Nonuniformity of gray distribution, correlation of gray gradient symbiosis matrix

GLCM, Gray-level co-occurrence matrix; *GLGCM*, gray-level–gradient co-occurrence matrix; *GLRLM*, gray-level run length matrix.

Deep learning methods comparison results

Four convolutional neural network models were used for classification experiments, namely, AlexNet, VGG16, VGG19, and Google Inception V3, and their network depths gradually deepened. The results of the transfer learning training network based on the Google Inception V3 model are the best. The average

Table III. Comparison of classification effects of melanomas by different types of instruments under the same conditions

Classifier	Average accuracy, %	Sensitivity, %	Specificity, %
SVM	81.21	71.04	91.23
KNN	76.73	77.12	76.20
LR	81.85	76.36	87.04
CART	75.44	61.42	89.11

CART, Classification and regression tree; *KNN*, k-nearest neighbors; *LR*, logistic regression; *SVM*, support vector machine.

Table IV. Result of deep learning models classification

Model	Training time, h:min	Average accuracy, %	Sensitivity, %	Specificity, %
AlexNet	1:15	72.20	64.50	79.90
VGG16	2:01	74.30	75.70	72.90
VGG19	2:13	76.60	65.70	77.50
Inception V3	2:37	93.70	95.30	92.10

VGG16, Visual Geometry Group Network 16; *VGG19*, Visual Geometry Group Network 19.

accuracy rate is close to 94%, and the sensitivity and specificity are 95.30% and 92.10%, respectively, followed by VGG16 and VGG19 with average accuracy rates of 74.30% and 76.60%, respectively. The training results are provided in Table IV.

As the depth of the network increases, the accuracy of the experimental classification also increases, and the time consumed also increases. In summary, Google Inception V3 is a convolutional neural network that is worthwhile to choose.

DISCUSSION

Comparison of traditional machine learning and deep learning

According to the above experimental results, the region growing segmentation algorithm incorporated with the LR classifier can achieve a better performance compared with the other algorithms in this review. In addition, color and texture features play a prominent role in distinguishing melanoma with other kinds of pigment, which were consistent with the clinical diagnostics of melanoma. Traditional machine learning was relatively satisfactory on small data sets, but it has low generation ability and requires the design of feature extractors for different problems. Because its feature extractor is designed for specific tasks, it has strong explanatory power but a high cost, which limits its use.

Table V. Advantages and disadvantages of traditional machine learning and deep learning

Traditional machine learning	Deep learning
Advantages	Advantages
Low data requirements	Can integrate different data sets;
The model is simple and easy to train	Feature extraction feature, no need to design feature extractor
Disadvantages	Disadvantages
Low generalization ability	Data-driven, large data requirements
High cost	High data quality requirements

Deep learning is a powerful tool that complements traditional machine learning. It is driven by data, and feature extraction is included, which has good generalization, but it requires a very large amount of training data (Table V). When the model is relatively deep, it is difficult to train and optimize the parameters and provide an explanation to clinical application after training.

Limitations

The melanoma samples used in this experiment were not divided by severity, and all of the non-melanoma samples were benign nevi. Further study should be made to classify the severity of melanoma, and the data used should be expanded to include some difficult cases such as dysplastic nevus. The small amount of data used in the experiment resulted in machine learning performing better than deep learning. However, with the increase of data, the advantages of deep learning will gradually become prominent. More data should be collected and used for further experiments.

Future outlook

The existing knowledge of dermatologists is invaluable for the diagnosis of melanoma. Because of the limited amount of medical data and various quality issues, incorporating expert knowledge into the deep learning process is an important research topic. At the same time, reliable content, such as Wikipedia and PubMed, can also be integrated to improve the overall performance of the system. Corresponding convolutional neural networks can be deployed in smartphones,[21] which will make them easy to use and help dermatologists in underdeveloped areas diagnose.

REFERENCES

1. Tripp MK, Watson M, Balk SJ, et al. State of the science on prevention and screening to reduce melanoma incidence and mortality: the time is now. *CA Cancer J Clin*. 2016;66(6):460-480.
2. Rigel DS, Carucci JA. Malignant melanoma: prevention, early detection, and treatment in the 21st century. *CA Cancer J Clin*. 2010;50(4):215-236.
3. Betta G, Leo GD, Fabbrocini G, et al. Automated application of the ''7-point checklist'' diagnosis method for skin lesions: estimation of chromatic and shape parameters. Presented at: the 2005 IEEE Instrumentation and Measurement Technology Conference, Ottawa, Ontario, Canada, May 17-19, 2005.
4. Dreiseitl S, Ohno-Machado L, Kittler H, et al. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *J Biomed Inform*. 2001;34(1):28-36.
5. Rahman Z, Jobson DJ, Woodell G, et al. Image enhancement, image quality, and noise. *Proc SPIE*. 2005;5907:59070N-59070N-15.
6. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. *Int J Comput Vis*. 1988;1(4):321-331.
7. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 2009;1(2):224-227.
8. Adams R, Bischof L. Seeded region growing. *IEEE Trans Pattern Anal Mach Intell*. 1994;16(6):641-647.
9. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybernetics*. 2007;9(1):62-66.
10. Stricker MA, Orengo M. Similarity of Color Images[J]. *Proc Spie Storage & Retrieval for Image & Video Databases*. 1995;2420:381-392.
11. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446.
12. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc*. 1996;58(1):267-288.
13. Adankon MM, Cheriet M. Support vector machine. *Comput Sci*. 2002;1(4):1-28.
14. Stone CJ. Classification and regression trees[J]. *Wadsworth International Group*. 1984;8:452-456.
15. Denœux T. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans Syst Man Cybernetics*. 1995;25(5):804-813.
16. Cucchiara A. Applied Logistic Regression[J]. *Technometrics*. 2012;34(3):358-359.
17. Pan SJ, Yang Q. A Survey on transfer learning[J]. *IEEE Transactions on Knowledge and Data Engineering*. 2010;22(10):1345-1359.
18. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottu L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 25*. Red Hook, NY: Curran Associates, Inc; 2012:1097-1105.
19. Simonyan K , Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition[J]*. Computer Science, 2014. Available at: https://arxiv.org/abs/1409.1556. Accessed August 26, 2019.
20. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2818-2826.
21. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.