Artificial Intelligence vs Medical Providers in the Dermoscopic Diagnosis of Melanoma

Jane M. Anderson, BSA; Izhaar Tejani, BA; Tory Jarmain, BA; Lisa Kellett, MD; Ronald L. Moy, MD

PRACTICE POINTS
- Artificial intelligence (AI) has the potential to facilitate the diagnosis of pigmented lesions and expedite the management of malignant melanoma.
- Further studies should be done to see if the high diagnostic accuracy of the AI application we studied translates to a decrease in unnecessary biopsies or expedited referral for pigmented lesions.
- The large variability of formal dermoscopy training among board-certified dermatologists may contribute to the decreased ability to identify pigmented lesions with dermoscopic imaging compared to AI.

Early diagnosis of melanoma drastically reduces morbidity and mortality; however, most skin lesions are not initially evaluated by dermatologists, and some patients may require a referral. This study sought to determine the performance of an artificial intelligence (AI) application in classifying lesions as benign or malignant to determine whether AI could assist in screening potential melanoma cases. One hundred dermoscopic images (80 benign nevi and 20 biopsy-verified malignant melanomas) were assessed by an AI application as well as 23 dermatologists, 7 family physicians, and 12 primary care mid-level providers. The AI's high accuracy and positive predictive value (PPV) demonstrate that this AI application could be a reliable melanoma screening tool for providers.

Cutis. 2023;111:254-258.

The incidence of skin cancer continues to increase, and it is by far the most common malignancy in the United States. Based on the sheer incidence and prevalence of skin cancer, early detection and treatment are critical. Looking at melanoma alone, the 5-year survival rate is greater than 99% when detected early but falls to 71% when the disease reaches the lymph nodes and 32% with metastasis to distant organs.[1] Furthermore, a 2018 study found stage I melanoma patients who were treated 4 months after biopsy had a 41% increased risk of death compared with those treated within the first month.[2] However, many patients are not seen by a dermatologist first for examination of suspicious skin lesions and instead are referred by a general practitioner or primary care mid-level provider. Therefore, many patients experience a longer time to diagnosis or treatment, which directly correlates with survival rate.

Dermoscopy is a noninvasive diagnostic tool for skin lesions, including melanoma. Using a handheld dermoscope (or dermatoscope), a transilluminating light source magnifies skin lesions and allows for the visualization of subsurface skin structures within the epidermis, dermo-epidermal junction, and papillary dermis.[3] Dermoscopy has been shown to improve a dermatologist's accuracy in diagnosing malignant melanoma vs clinical evaluation with the unaided eye.[4,5] More recently, dermoscopy has been digitized, allowing for the collection and documentation of case photographs. Dermoscopy also has expanded past the scope of dermatologists and has become increasingly useful in primary care.[6] Among family physicians, dermoscopy also has been shown to have a higher sensitivity for melanoma detection compared to gross examination.[7] Therefore, both the increased diagnostic performance of malignant melanoma using a dermoscope and the expanded use of dermoscopy in medical care validate the evaluation of an artificial intelligence (AI)

algorithm in diagnosing malignant melanoma using dermoscopic images.

Triage (Triage Technologies Inc) is an AI application that uses a web interface and combines a pretrained convolutional neural network (CNN) with a reinforcement learning agent as a question-answering model. The CNN algorithm can classify 133 different skin diseases, 7 of which it is able to classify using dermoscopic images. This study sought to evaluate the performance of Triage's dermoscopic classifier in identifying lesions as benign or malignant to determine whether AI could assist in the triage of skin cancer cases to shorten time to diagnosis.

Materials and Methods

The MClass-D test set from the International Skin Imaging Collaboration was assessed by both AI and practicing medical providers. The set was composed of 80 benign nevi and 20 biopsy-verified malignant melanomas. Board-certified US dermatologists (n=23), family physicians (n=7), and primary care mid-level providers (n=12)(ie, nurse practitioners, physician assistants) were asked to label the images as benign or malignant. The results from the medical providers were then compared to the performance of the AI application by looking at the sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). Statistical significance was determined with a 1 sample t test run through RStudio (Posit Software, PBC), and $P<.05$ was considered significant.

Results

The AI application performed extremely well in differentiating between benign nevi and malignant melanomas, with a sensitivity of 80%, specificity of 95%, accuracy of 92%, PPV of 80%, and NPV of 95% (Table 1). When compared with practicing medical providers, the AI performed significantly better in almost all categories ($P<.05$)(Figure 1). With all medical providers combined, the AI had significantly higher accuracy, sensitivity, and specificity ($P<.05$). The accuracy of the individual medical providers ranged from 32% to 78%.

Compared with dermatologists, the AI was significantly more specific and accurate and demonstrated a higher PPV and NPV ($P<.05$). There was no significant difference between the AI and dermatologists in sensitivity or labeling the true malignant lesions as malignant. The dermatologists who participated had been practicing from 1.5 years to 44 years, with an average of 16 years of dermatologic experience. There was no correlation between years practicing and performance in determining the malignancy of lesions. Of 14 dermatologists, dermoscopy was used daily by 10 and occasionally by 3, but only 6 dermatologists had any formal training. Dermatologists who used dermoscopy averaged 11 years of use.

The AI also performed significantly better than the primary care providers, including both family physicians and mid-level providers ($P<.05$). With the family physicians and mid-level provider scores combined, the AI showed a statistically significantly better performance in all categories examined, including sensitivity, specificity, accuracy, PPV, and NPV ($P<.05$). However, when compared with family physicians alone, the AI did not demonstrate a statistically significant difference in sensitivity.

Comment

Automatic Visual Recognition Development—The AI application we studied was developed by dermatologists as a tool
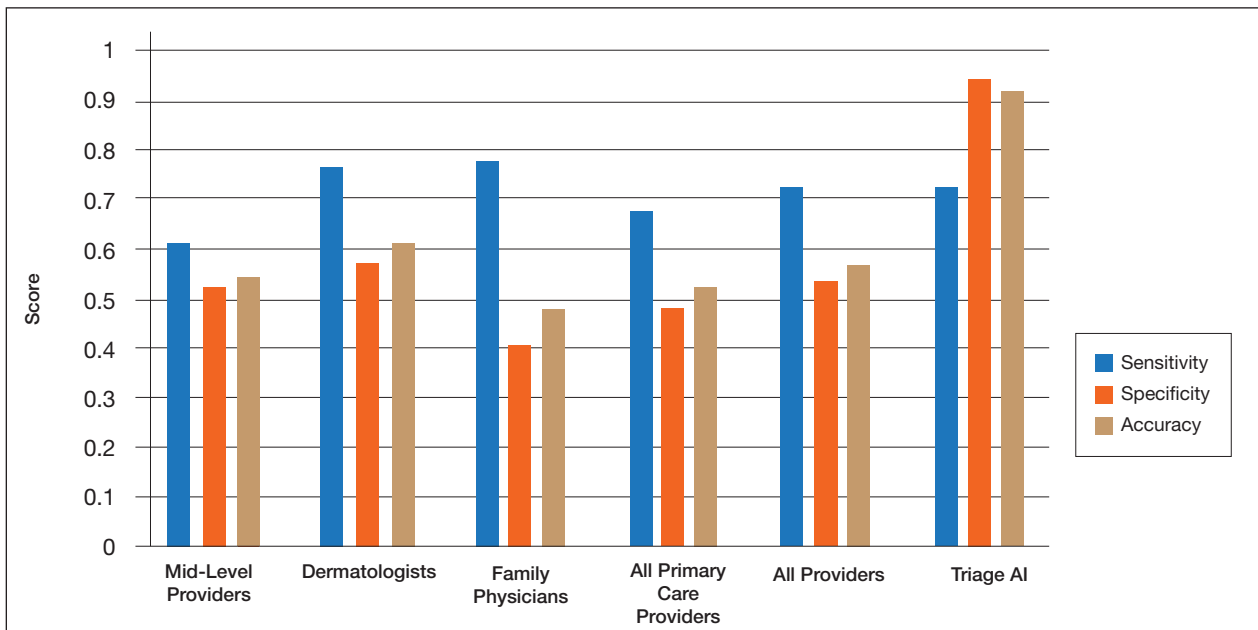

FIGURE 1. Performance of artificial intelligence (AI)(Triage Technologies Inc) vs medical providers in differentiating benign nevi vs malignant melanoma.

to assist in the screening of skin lesions suspicious for melanoma or a benign neoplasm.[8] Developing AI applications that can reliably recognize objects in photographs has been the subject of considerable research. Notable progress in automatic visual recognition was shown in 2012 when a deep learning model won the ImageNet object recognition challenge and outperformed competing approaches by a large margin.[9,10] The ImageNet competition, which has been held annually since 2010, required participants to build a visual classification system that distinguished among 1000 object categories using 1.2 million labeled images as training data. In 2017, participants developed automated visual systems that surpassed the estimated human performance.[11] Given this success,

the organization decided to deliver a more challenging competition involving 3D imaging—Medical ImageNet, a petabyte-scale, cloud-based, open repository project—with goals including image classification and annotation.[12]

Convolutional Neural Networks—Convolutional neural networks are computer system architectures commonly employed for making predictions from images.[13] Convolutional neural networks are based on a set of layers of learned filters that perform convolution, a mathematical operation that reflects the relationship between the 2 functions. The main algorithm that makes the learning possible is called *backpropagation*, wherein an error is computed at the output and distributed backward through the neural network's layers.[14] Although

TABLE 1. Performance of the AI Application Compared With Practicing Medical Providers[a]

Category of provider	Sensitivity	Specificity	Accuracy	PPV	NPV
All primary care providers (n=19)	.67[b]	.48[b]	.52[b]	.25[b]	.87[b]
Family physicians (n=7)	.78	.41[b]	.48[b]	.25[b]	.90[b]
Mid-level providers (PAs and NPs)(n=12)	.61[b]	.53[b]	.55[b]	.26[b]	.85[b]
Dermatologists (n=23)	.77	.57[b]	.61[b]	.32[b]	.92[b]
Triage AI	.80	.95	.92	.80	.95

Abbreviations: AI, artificial intelligence; NP, nurse practitioner; NPV, negative predictive value; PA, physician assistant; PPV, positive predictive value.

[a]Triage (Triage Technologies Inc).

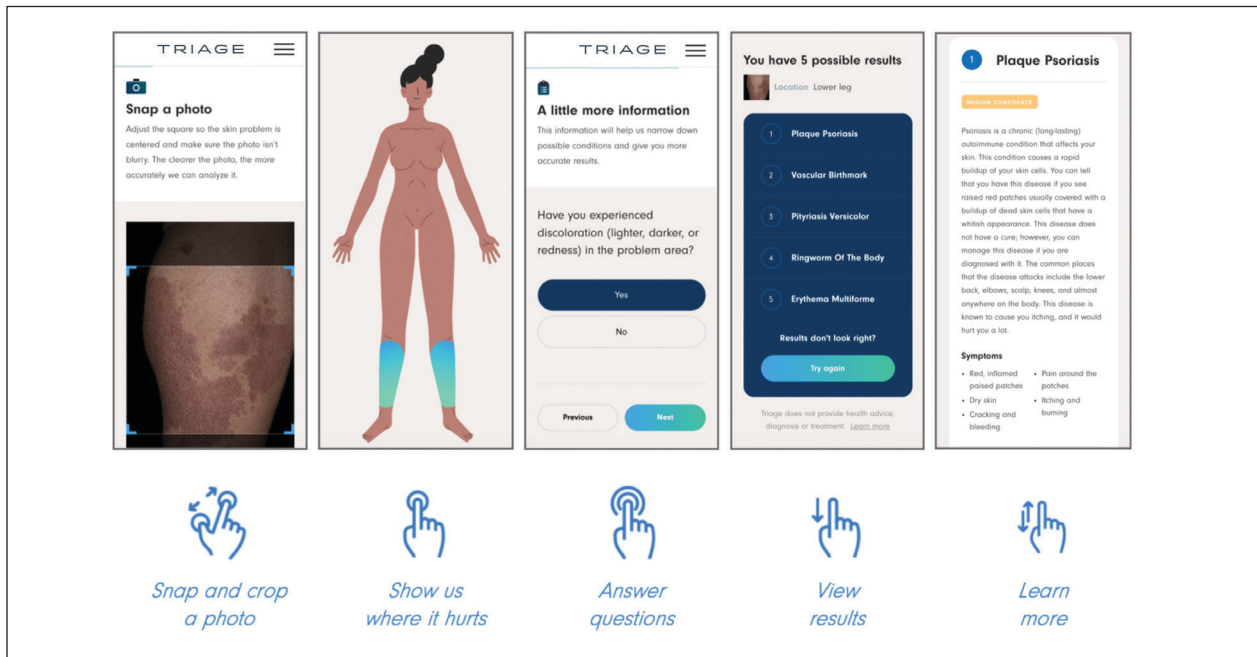[b]$P<.05$ when compared with the AI application.

FIGURE 2. Artificial intelligence application interface. Image courtesy of Triage Technologies Inc and Izhaar Tejani, BA (Toronto, Ontario, Canada).

CNNs and backpropagation methods have existed since 1989, recent technologic advances have allowed for deep learning–based algorithms to be widely integrated with everyday applications.[15] Advances in computational power in the form of graphics processing units and parallelization, the existence of large data sets such as the ImageNet database, and the rise of software frameworks have allowed for quick prototyping and deployment of deep learning models.[16,17]

Convolutional neural networks have demonstrated potential to excel at a wide range of visual tasks. In dermatology, visual recognition methods often rely on using either a pretrained CNN as a feature extractor for further classification or fine-tuning a pretrained network on dermoscopic images.[18-20] In 2017, a model was trained on 130,000 clinical images of benign and malignant skin lesions. Its performance was found to be in line with that of 21 US board-certified dermatology experts when diagnosing skin cancers from clinical images confirmed by biopsy.[21]

Triage—The AI application Triage is composed of several components contained in a web interface (Figure 2). To use the interface, the user must sign up and upload a photograph to the website. The image first passes through a gated-logic visual classifier that rejects any images that do not contain a visible skin condition. If the image contains a skin condition, the image is passed to a skin classifier that predicts the probability of the image containing 1 of 133 classes of skin conditions, 7 of which the application can diagnose with a dermoscopic image.

The AI application uses several techniques when training a CNN model. To address skin condition class imbalances (when more examples exist for 1 class than the others) in the training data, additional weights are applied to mistakes made on underrepresented classes, which encourages the model to better detect cases with low prevalence in the data set. Data augmentation techniques such as rotating, zooming, and flipping the training images are applied to allow the model to become more familiar with variability in the input images. Convolutional neural networks are trained using a well-known neural network optimization method called *Stochastic gradient descent with momentum*.[22]

The final predictions are refined by a question-and-answer system that encodes dermatology knowledge and is currently under active development. Finally, the top k most probable conditions are displayed to the user, where k≤5. An initial prototype of the system was described in a published research paper in the 2019 medical imaging workshop of the Neural Information Systems conference.[23]

The prototype demonstrated that combining a pretrained CNN with a reinforcement learning agent as a question-answering model increased the classification confidence and accuracy of its visual symptom checker and decreased the average number of questions asked to narrow down the differential diagnosis. The reinforcement learning approach increases the accuracy more than 20% compared with the CNN-only approach, which only uses visual information to predict the condition.[23]

This application's current visual question-answering system is trained on a diverse set of data that includes more than 20 years of clinical encounters and user-uploaded cases submitted by more than 150,000 patients and 10,000 clinicians in more than 150 countries. All crowdsourced images used for training the dermoscopy classifier are biopsy-verified images contributed by dermatologists. These data are made up of case photographs that are tagged with metadata around the patient's age, sex, symptoms, and diagnoses. The CNN algorithm used covers 133 skin disease classes, representing 588 clinical conditions. It also can automatically detect 7 malignant, premalignant, and benign dermoscopic categories, which is the focus of this study (Table 2). Diagnoses are verified by patient response to treatment, biopsy results, and dermatologist consensus.

In addition to having improved performance, supporting more than 130 disease classes, and having a diverse data set, the application used has beat competing technologies.[20,24] The application currently is available on the internet in more than 30 countries after it received Health Canada Class I medical device approval and the CE mark in Europe.

Can AI Reliably Detect Melanoma?—In our study, of the lesions labeled benign, the higher PPV and NPV of the AI algorithm means that the lesions were more reliably true benign lesions, and the lesions labeled as malignant were more likely to be true malignant lesions. Therefore, the diagnosis given by the AI compared with the medical provider was significantly more likely to be correct. These findings demonstrate that this AI application can reliably detect malignant melanoma using dermoscopic images. However, this study was limited by the small sample size of medical providers. Further studies are necessary to

TABLE 2. Dermoscopic Disease Categories Supported by an Artificial Intelligence Application[8,a]

Disease category	Malignancy
Actinic keratosis	Premalignant
Basal cell carcinoma	Malignant
Benign melanocytic neoplasm	Benign
Dermatofibroma	Benign
Melanoma	Malignant
Seborrheic keratosis	Benign
Squamous cell carcinoma	Malignant

aTriage (Triage Technologies Inc).

assess whether the high diagnostic accuracy of the application translates to expedited referrals and a decrease in unnecessary biopsies.

Dermoscopy Training—This study looked at dermoscopic images instead of gross examination, as is often done in clinic, which draws into question the dermoscopic training dermatologists receive. The diagnostic accuracy using dermoscopic images has been shown to be higher than evaluation with the naked eye.[5,6] However, there currently is no standard for dermoscopic training in dermatology residencies, and education varies widely.[25] These data suggest that there may be a lack of dermoscopic training among dermatologists, which could accentuate the difference in performance between dermatologists and AI. Most primary care providers also lack formal dermoscopy training. Although dermoscopy has been shown to increase the diagnostic efficacy of primary care providers, this increase does not become apparent until the medical provider has had years of formal training in addition to clinical experience, which is not commonly provided in the medical training that primary care providers receive.[8,26]

Conclusion

It is anticipated that AI will shape the future of medicine and become incorporated into daily practice.[27] Artificial intelligence will not replace physicians but rather assist clinicians and help to streamline medical care. Clinicians will take on the role of interpreting AI output and integrate it into patient care. With this advancement, it is important to highlight that for AI to improve the quality, efficiency, and accessibility of health care, clinicians must be equipped with the right training.[27-29]

REFERENCES

1. Cancer facts & figures 2023. American Cancer Society. Accessed April 20, 2023. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf
2. Conic RZ, Cabrera CI, Khorana AA, et al. Determination of the impact of melanoma surgical timing on survival using the National Cancer Database. *J Am Acad Dermatol*. 2018;78:40-46.e7. doi:10.1016/j.jaad.2017.08.039
3. Lallas A, Zalaudek I, Argenziano G, et al. Dermoscopy in general dermatology. *Dermatol Clin*. 2013;31:679-694, x. doi:10.1016/j.det.2013.06.008
4. Bafounta M-L, Beauchet A, Aegerter P, et al. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma?: results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. *Arch Dermatol*. 2001;137:1343-1350. doi:10.1001/archderm.137.10.1343
5. Vestergaard ME, Macaskill P, Holt PE, et al. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *Br J Dermatol*. 2008;159:669-676. doi:10.1111/j.1365-2133.2008.08713.x
6. Marghoob AA, Usatine RP, Jaimes N. Dermoscopy for the family physician. *Am Fam Physician*. 2013;88:441-450.
7. Herschorn A. Dermoscopy for melanoma detection in family practice. *Can Fam Physician*. 2012;58:740-745, e372-8.
8. Instructions for use for the Triage app. Triage website. Accessed April 20, 2023. https://www.triage.com/pdf/en/Instructions%20for%20Use.pdf
9. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, et al, eds. *Advances in Neural Information Processing Systems*. Vol 25. Curran Associates, Inc; 2012. Accessed April 17, 2023. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
10. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115:211-252. doi:10.1007/s11263-015-0816-y
11. Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks. *IEEE Trans Patt Anal Mach Intell*. 2020;42:2011-2023. doi:10.1109/TPAMI.2019.2913372
12. Medical image net-radiology informatics. Stanford University Center for Artificial Intelligence in Medicine & Imaging website. Accessed April 20, 2023. https://aimi.stanford.edu/medical-imagenet
13. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444. doi:10.1038/nature14539
14. Le Cun Y et al. A theoretical framework for back-propagation. In: Touretzky D, Honton G, Sejnowski T, eds. *Proceedings of the 1988 Connect Models Summer School*. Morgan Kaufmann; 1988:21-28.
15. Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86:2278-2324. doi:10.1109/5.726791
16. Chollet E. About Keras. Keras website. Accessed April 21, 2023. https://keras.io/about/
17. Introduction to TensorFlow. TensorFlow website. Accessed April 21, 2023. https://www.tensorflow.org/learn
18. Kawahara J, BenTaieb A, Hamarneh G. Deep features to classify skin lesions. *2016 IEEE 13th International Symposium on Biomedical Imaging*. 2016. doi:10.1109/ISBI.2016.7493528
19. Lopez AR, Giro-i-Nieto X, Burdick J, et al. Skin lesion classification from dermoscopic images using deep learning techniques. doi:10.2316/P.2017.852-053
20. Codella NCF, Nguyen QB, Pankanti S, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM J Res Dev*. 2017;61:1-28. doi:10.1147/JRD.2017.2708299
21. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118. doi:10.1038/nature21056
22. Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning. *ICML'13: Proceedings of the 30th International Conference on International Conference on Machine Learning*. 2013;28:1139-1147.
23. Akrout M, Farahmand AM, Jarmain T, et al. Improving skin condition classification with a visual symptom checker trained using reinforcement learning. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019: 22nd International Conference*. October 13-17, 2019. Shenzhen, China. Proceedings, Part IV. Springer-Verlag; 549-557. doi:10.1007/978-3-030-32251-9_60
24. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26:900-908. doi:10.1038/s41591-020-0842-3
25. Fried LJ, Tan A, Berry EG, et al. Dermoscopy proficiency expectations for US dermatology resident physicians: results of a modified delphi survey of pigmented lesion experts. *JAMA Dermatol*. 2021;157:189-197. doi:10.1001/jamadermatol.2020.5213
26. Fee JA, McGrady FP, Rosendahl C, et al. Training primary care physicians in dermoscopy for skin cancer detection: a scoping review. *J Cancer Educ*. 2020;35:643-650. doi:10.1007/s13187-019-01647-7
27. James CA, Wachter RM, Woolliscroft JO. Preparing clinicians for a clinical world influenced by artificial intelligence. *JAMA*. 2022;327:1333-1334. doi:10.1001/jama.2022.3580
28. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2:719-731. doi:10.1038/s41551-018-0305-z
29. Chen M, Decary M. Artificial intelligence in healthcare: an essential guide for health leaders. *Healthc Manag Forum*. 2020;33:10-18. doi:10.1177/0840470419873123