Artificial Intelligence in Dermatology: A Primer

Albert T. Young[1,2], Mulin Xiong[3], Jacob Pfau[1,2], Michael J. Keiser[4] and Maria L. Wei[1,2,5]

Artificial intelligence is becoming increasingly important in dermatology, with studies reporting accuracy matching or exceeding dermatologists for the diagnosis of skin lesions from clinical and dermoscopic images. However, real-world clinical validation is currently lacking. We review dermatological applications of deep learning, the leading artificial intelligence technology for image analysis, and discuss its current capabilities, potential failure modes, and challenges surrounding performance assessment and interpretability. We address the following three primary applications: (i) teledermatology, including triage for referral to dermatologists; (ii) augmenting clinical assessment during face-to-face visits; and (iii) dermatopathology. We discuss equity and ethical issues related to future clinical adoption and recommend specific standardization of metrics for reporting model performance.

Journal of Investigative Dermatology (2020) ■, ■−■; doi:10.1016/j.jid.2020.02.026

INTRODUCTION

Artificial intelligence (AI) is transforming health care (Naylor, 2018). Deep learning (DL) has become the dominant AI technology for high-dimensional complex data, such as images (Esteva et al., 2019). In brief, DL leverages artificial neural networks, which learn complex mappings between inputs (e.g., images) and outputs (e.g., diagnoses) without explicit human engineering. Inspired by the brain, artificial neurons arranged in deep layers adapt the strength of their connections to one another as the model self-learns features from the input, such as visual patterns, that are most relevant for predicting the output.

In experimental settings across multiple specialties, DL performs equivalently to health-care professionals for detecting disease from medical imaging (Liu et al., 2019a). An AI system (Moleanalyzer Pro) has been approved for the European market as a medical device and has been shown to perform comparably with dermatologists in a setting simulating store-and-forward dermatology (Haenssle et al., 2020).

Dermatology is well positioned to leverage DL to improve patient care, with its emphasis on visual analysis. Given the shortage of dermatologists in the U.S. (Jayakumar and Lipoff, 2019) and the increasing incidence of cutaneous melanoma (National Cancer Institute, 2020), AI may play an increasingly important role in improving access to and quality of dermatological care. This review summarizes research on the automated classification and monitoring of skin lesions, discusses barriers to clinical adoption, and proposes metrics for AI model performance.

Where we are now

AI research in dermatology initially focused on skin cancer, particularly melanoma; more recently, it has taken on multiple classes of diagnoses and therapeutic recommendations. A meta-analysis of 70 studies found the accuracy of computer-aided diagnosis of melanoma to be comparable to that of human experts (Dick et al., 2019); using convolutional neural networks (CNNs), the leading DL algorithm for image analysis, many studies have reported dermatologist-level classification of cutaneous lesions from dermoscopic and nondermoscopic images (Table 1).

Nondermoscopic images. A CNN trained on 129,450 images achieved performance comparable to dermatologists on two binary classification tasks, carcinomas versus seborrheic keratoses and melanomas versus nevi, for both dermoscopic and nondermoscopic images (Esteva et al., 2017). Subsequently, dermatologist-level classification of malignant versus benign lesions using nondermoscopic datasets of predominantly East Asian participants was achieved (Fujisawa et al., 2019; Han et al., 2019a, 2018a). Han et al. (2019b) reported an area under the receiver operating characteristic curve of 0.94 for malignancy detection among 134 disorders, on par with dermatology residents; they also reported an area under the curve of 0.89−0.94 for predicting appropriate medications among four primary treatment options. CNNs have also classified onychomycosis (Han et al., 2018b) and lip diseases at a level similar to dermatologists (Cho et al., 2019).

Dermoscopic images. CNNs have classified dermoscopic images of melanoma versus nevi with performances similar to or exceeding dermatologists (Brinker et al., 2019a, 2019b; Codella et al., 2016; Haenssle et al., 2020, 2018; Hekler et al., 2019a; Marchetti et al., 2018, 2020; Phillips et al., 2019; Tschandl et al., 2019b). CNNs have also achieved expert-level diagnosis of nonpigmented skin cancer (Tschandl et al., 2019c) and outperformed dermatologists across five disease classes (Maron et al., 2019). Switching imaging modalities, a CNN trained only on dermoscopic

[1] Department of Dermatology, University of California, San Francisco, San Francisco, California, USA; [2] Dermatology Service, San Francisco Veterans Affairs Medical Center, San Francisco, California, USA; [3] Michigan State University College of Human Medicine, East Lansing, Michigan, USA; [4] Department of Pharmaceutical Chemistry, Department of Bioengineering and Therapeutic Sciences, Institute for Neurodegenerative Diseases, and Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California, USA; and [5] Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California, USA

Correspondence: Maria L. Wei, San Francisco Veterans Affairs Medical Center, 1700 Owens Street, Room 324, San Francisco, California, 94158. E-mail: maria.wei@ucsf.edu

Abbreviations: AI, artificial intelligence; CNN, convolutional neural network; DL, deep learning

Received 3 December 2019; revised 22 February 2020; accepted 25 February 2020; accepted XXX; corrected proof published online XXX

Table 1. Dermatologist-Level Classification of Skin Lesions with Convolutional Neural Networks

Study	Location	Dataset	Classification Task	Algorithm Performance[1]	Clinician Performance[2]
Dermoscopic and nondermoscopic test images: Binary classification					
Esteva et al. (2017)	USA	129,450 clinical images, including 3,374 dermoscopic images of 757 disease classes	Binary: (1) Keratinocyte carcinoma versus SK; (2) melanomas versus nevi	AUC 0.96 (nondermoscopic images) AUC 0.91 (dermoscopic images)	Comparable[3] sensitivity and specificity, 21 board-certified dermatologists
Nondermoscopic test images: Binary classification					
Han et al. (2018b)	South Korea	Training set: 49,567 images Test set: 1,358 images	Binary: Onychomycosis or not	AUC 0.82–0.98 for diagnosis of onychomycosis, depending on validation dataset	Comparable[3] sensitivity and specificity, 42 dermatologists
Brinker et al. (2019c)	Germany	12,378 open-source dermoscopic images	Binary: Melanoma versus atypical nevi	89.4% sensitivity[4] and 68.2% specificity	89.4% sensitivity and 64.4% specificity by 145 dermatologists of all levels of training from 12 German university hospitals
Cho et al. (2019)	South Korea	Training set: 1,629 images (743 malignant and 886 benign). Test set: 625 images (from two other hospitals)	Binary: Malignant versus benign lip disorders	AUC 0.81 for diagnosis of lip malignancy	Comparable[3] sensitivity and specificity, 44 participants (6 board-certified dermatologists, 12 dermatology residents, 9 medical doctors not specialized in dermatology, and 17 medical students)
Fujisawa et al. (2019)	Japan	6,009 images (4,867 train and 1,142 test) of 14 diagnoses, including malignant and benign conditions	Binary: Benign versus malignant lesions	92.4% accuracy 96.3% sensitivity and 89.5% specificity	85.3% accuracy by 13 board-certified dermatologists 74.4% accuracy by 9 dermatology trainees
Han et al. (2019a)	South Korea	Training set: 182,348 clinical images Test set: 2,844 images	Binary: Benign versus malignant lesions	AUC 0.919	ROC area 0.906
Nondermoscopic test images: Multiclass classification					
Han et al. (2018a)	South Korea	182,014 clinical images	Multiclass: 12 disease classes (BCC, SCC, intraepithelial carcinoma, AK, SK, melanocytic nevus, lentigo, dermatofibroma, pyogenic granuloma, hemangioma, and wart)	AUC 0.96 on Asan dataset (Asian) AUC 0.88 on Edinburgh dataset (Caucasian)	Comparable[3] sensitivity and specificity, 16 dermatologists (10 professors and 6 clinicians)
Liu et al. (2019)	USA	Adult cases from a teledermatology service serving two states in the U.S. Training set: 14,021 cases Test set: 3,756 cases	Multiclass: 26 disease classes (common skin conditions, representing roughly 80% of the volume of skin conditions seen in a primary care setting)	0.67 top-1 accuracy and 0.90 top-3 accuracy over 26 diagnoses	0.63 top-1 accuracy and 0.75 top-3 accuracy, 6 board-certified dermatologists
Dermoscopic test images: Binary classification					
Codella et al. (2016)	USA	1,279 images (900 train and 379 test)	Binary: Melanoma versus melanocytic nevi	82% sensitivity[4] and 62% specificity AUC 0.84	82% sensitivity and 59% specificity, 8 expert dermatologists
Marchetti et al. (2018)	Multiple countries	1,279 images (900 train and 379 test)	Binary: Melanoma versus melanocytic nevi	82% sensitivity[4] and 76% specificity AUC 0.86	82% sensitivity and 59% specificity, 8 expert dermatologists from four countries
Haenssle et al. (2018)	Germany	>100,000 dermoscopic images	Binary: Melanoma versus benign melanocytic nevi	AUC 0.86 (more difficult test-set-100); AUC 0.95 (test-set-300)	AUC 0.79, international group of 58 dermatologists

Table 1. Continued

Study	Location	Dataset	Classification Task	Algorithm Performance[1]	Clinician Performance[2]
Brinker et al. (2019b)	Germany	12,378 open-source dermoscopic images	Binary: Melanoma versus atypical nevi	74.1% sensitivity[4] and 86.5% specificity	74.1% sensitivity and 60% specificity, 157 dermatologists of all levels of training from 12 German university hospitals
Hekler et al. (2019b)	Germany	Training set: 4,204 biopsy-proven images of melanoma and nevi (1:1) Test set: 804 biopsy-proven images of melanoma and nevi (1:1)	Binary: Melanoma versus nevi	82.3% sensitivity and 77.9% specificity	67.2% sensitivity and 62.2% specificity, dermatologists from 9 German university hospitals (each test image evaluated an average of 21.3 times)
Phillips et al. (2019)	United Kingdom	Training set: not reported Test set: 551 biopsied lesions (including 125 melanoma) and 999 control lesions (assumed benign)	Binary: Melanoma versus nonmelanoma	100% sensitivity[4] and 64.8% specificity with iPhone 6s images	100% sensitivity[5] and 69.9% specificity; no description of clinicians provided
Tschandl et al. (2019c)	Austria, Australia	Training set: 7,895 dermoscopic and 5,829 close-up images Test set: 2,072 dermoscopic and clinical close-up images	Binary: Malignant versus benign nonpigmented skin lesions	80.5% sensitivity and 51.3% specificity AUC 0.74	77.6% sensitivity and 51.3% specificity; AUC 0.70, 95 raters, including 62 board-certified dermatologists
Dermoscopic test images: Multiclass classification					
Marchetti et al. (2020)	USA	Training set: ∼2,000 images Test set: 150 images	Multiclass: 3 disease classes (SK, melanoma, and nevus)	76% sensitivity[4] and 85% specificity AUC 0.87	76.0% sensitivity, 72.6% specificity AUC 0.74
Maron et al. (2019)	Germany	Training set: 11,444 dermoscopic images Test set: 300 biopsy-verified images	Multiclass: 5 disease classes (AK, intraepithelial carcinoma, benign keratosis, melanocytic nevi, and melanoma)	AUC 0.96 macro-mean AUC for multiclass AUC 0.93 for benign versus malignant	112 dermatologists from 13 university hospitals; performance was below the model's average performance
Tschandl et al. (2019b)	Multiple countries	Training set: 10,015 dermoscopic images Test set: 1,195 images	Multiclass: 7 disease classes (intraepithelial carcinoma including AK and Bowen's disease; BCC; benign keratinocytic lesions including solar lentigo, SK, and LPLK; dermatofibroma; melanoma; melanocytic nevi; and vascular lesions)	81.9% sensitivity and 96.2% specificity (top three algorithms of 139 challenge submissions)	67.8% sensitivity and 94.0% specificity (by majority vote), 27 expert readers 73.1% sensitivity and 92.8% specificity (by majority vote), 511 readers, 63 countries (283 board-certified dermatologists, 118 dermatology residents, and 83 general practitioners)
Haenssle et al. (2020)	Multiple countries	Dermoscopic images from multiple sources	Multiclass: 10 disease classes (nevus, angioma/ angiokeratoma, SK, dermatofibroma, solar lentigo, AK, Bowen's disease, melanoma, BCC, and SCC)	95.0% sensitivity, 80.4% specificity[4] for benign versus malignant	94.1% sensitivity, 80.4% specificity by 96 dermatologists for management decision (given clinical close-up images, dermoscopy, and textual information)

Abbreviations: AK, actinic keratosis; AUC, area under the receiver operating characteristic curve; BCC, basal cell cancer; LPLK, lichen planus-like keratosis; ROC, receiver operating characteristic; SCC, squamous cell cancer; SK, seborrheic keratosis

[1] For diagnosis of melanoma, unless otherwise indicated.

[2] Compared with algorithm and calculated via mean, unless otherwise specified.

[3] Sensitivity and specificity were not directly reported, but most dermatologists fell below the algorithm's ROC curve.

[4] At threshold selected to match dermatologists' sensitivity or specificity.

[5] 100% sensitivity was guaranteed by the study design because there was no mechanism to detect false negatives.

images nonetheless achieved dermatologist-level melanoma classification performance on nondermoscopic images (Brinker et al., 2019c).

Alternative imaging modalities. AI coupled with hardware-based methods such as spectroscopy, multispectral imaging, or other specialized imaging modalities may augment dermatologists' capabilities (Dick et al., 2019; Ferrante di Ruffano et al., 2018; Szyc et al., 2019). For example, early melanomas may not present morphologic differences detectable by conventional photography, but computer-assisted techniques like dermatofluoroscopy may provide additional information for early diagnosis. Furthermore, the use of AI with these modalities obviates the need for specialized operator training.

Emerging applications

Teledermatology. Telemedicine may be one of the first fields to embrace AI, driven by demand for services, the necessity of collecting fit-for-purpose high-quality images, and the availability of existing technology (Xiong et al., 2019). Face-to-face diagnostic accuracy exceeds that of teledermatology (Finnane et al., 2017); however, inequalities surrounding access to dermatological care persist. Teledermatology has the potential to increase access by facilitating referrals and offering convenience and decreased wait times (Finnane et al., 2017), as well as providing diagnostic support at the time of case review. For teledermatology cases, the accuracy of a DL classifier (0.67) matched dermatologists' (0.63) and was higher than primary care physicians' (0.45) for 26 skin conditions (Liu et al., 2019b).

AI may be integrated into smartphone apps to photograph skin lesions, collect relevant clinical information, and generate a referral if appropriate. Many smartphones already support on-device DL with Google's TensorFlow Lite (TensorFlow, 2020) or Apple's CoreML (Apple Inc, 2020), preserving privacy by keeping health information on the device. A systematic review found nine studies that evaluated six algorithm-based smartphone apps and concluded that evidence of diagnostic accuracy was poor and does not support current implementation, despite two apps having obtained the CE marking; no apps are Food and Drug Administration approved (Freeman et al., 2020).

AI may also assist in automatic tracking and monitoring of skin lesions; although preliminary results are promising, existing studies used small datasets with little description, and there is no established standard metric of change (Navarro et al., 2019). Further study hinges on the prospective collection of large datasets.

Augmenting face-to-face assessments. AI may enhance care by providing diagnostic support in real-time during a clinical visit. Using clinical images, the top-1 and top-3 accuracies (indicating the fraction of cases where the top-*n* diagnoses contained the correct diagnosis) of dermatologists in diagnosing 134 skin disorders were increased by 7.0% and 10.1%, respectively, with AI (Han et al., 2019b). For dermoscopic images, the combination of AI and humans achieved an accuracy of 83.0% (compared with 81.6% and 42.9% achieved by AI and humans alone, respectively). About half of skin-related physician visits are to nondermatologists, who have variable training in diagnosing and managing skin conditions (Wilmer et al., 2014) and are less accurate than dermatologists in diagnosing melanoma (Martinka et al., 2016); AI-assisted diagnosis will likely have an even greater benefit for primary care physician skin exams.

AI can also expand physician differential diagnoses by retrieving images from a reference library with the most similar features to a concerning lesion (VisualDx, 2020); further study is needed to assess the efficacy of such systems.

It is unknown how CNNs perform compared with dermatologists making face-to-face assessments because studies report dermatologist-level diagnostic accuracy based on clinician evaluations of images in an artificial setting, using curated images, and without providing the full complement of meta-data normally available in clinic and teledermatology settings. Dermatologists improved their diagnostic accuracy when given access to close-up images and limited clinical information such as age, sex, and body site (Haenssle et al., 2020).

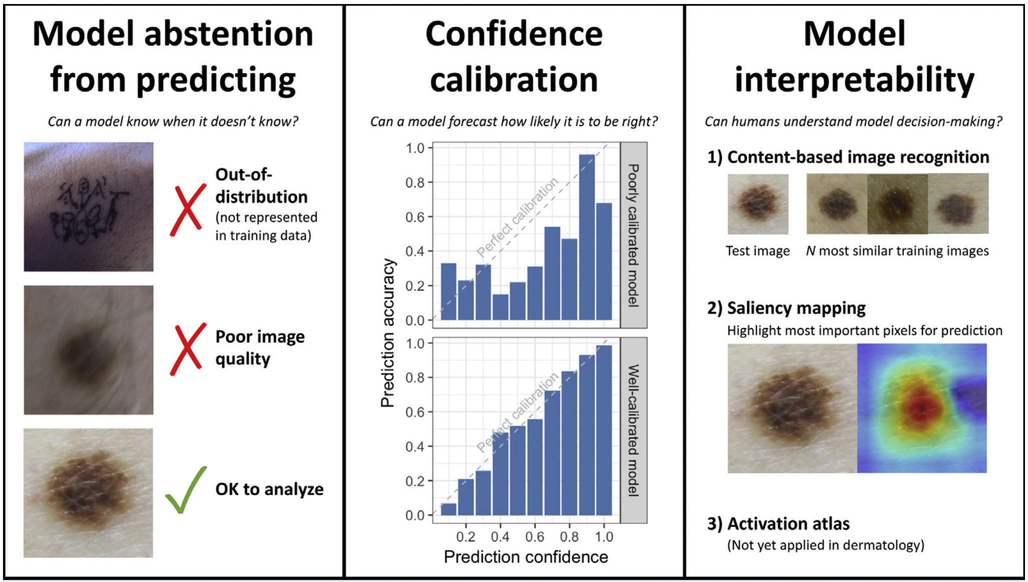
Dermatopathology

Histopathology is the gold standard for skin lesion diagnosis, but studies have shown poor inter- and intra-rater concordance and reproducibility for melanoma diagnosis (Piepkorn et al., 2019). AI has the potential to increase the accuracy and reproducibility of results, particularly if molecular diagnostics are used for model training. AI-augmented dermatopathology may also increase access to evaluation in areas where dermatopathologists are scarce. Evidence supports slide digitization; diagnosis on scanned cutaneous whole-slide images has comparable accuracy and reproducibility to diagnosis on glass slides (Onega et al., 2018).

DL has achieved clinical-grade performance on histopathologic classification of basal cell carcinoma, prostate cancer, and breast cancer metastases on whole-slide images (Campanella et al., 2019); outperformed pathologists on the classification of melanoma on cropped whole-slide images (Hekler et al., 2019b); and achieved an accuracy of 78% on an entire dermatopathology test set and an accuracy of 98% on the top 20% most confident predictions for classifying whole-slide images into one of four classes (Ianni et al., 2019). DL may help triage the most challenging cases, such as atypical melanocytic lesions, for focused review (Onega et al., 2018). Prospective studies are needed to assess the clinical impact of DL-assisted histopathologic diagnosis.

Considerations surrounding clinical adoption

Equity. AI has the potential to worsen health-care disparities, as recognized by the popular media (Khullar, 2019), particularly in dermatology (Adamson and Smith, 2018). The first concern is adequate representation of underserved populations in training data. Existing DL models have been trained on mainly European or East Asian populations, and the relative lack of training on darker skin pigmentation may limit overall diagnostic accuracy. This possibility is demonstrated by the increased error rates in commercial systems, trained on predominantly white datasets, for facial analysis in identifying black individuals (Buolamwini and Gebru, 2018). Second, AI may entrench existing social and economic biases


Model abstention from predicting
Can a model know when it doesn't know?
Out-of-distribution
(not represented in training data)
Poor image quality
OK to analyze
Confidence calibration
Can a model forecast how likely it is to be right?
Prediction accuracy
Prediction confidence
1.0
0.8
0.6
0.4
0.2
0.0
Perfect calibration
Poorly calibrated model
Well-calibrated model
Model interpretability
Can humans understand model decision-making?
1) Content-based image recognition
Test image
N most similar training images
2) Saliency mapping
Highlight most important pixels for prediction
3) Activation atlas
(Not yet applied in dermatology)

Recommended metric to report	Specific metrics
Dataset characteristics	- Demographics of train and test sets - Number of images per disease class
Model disease discrimination	- Area under the receiver operating characteristic curve (AUROC) - Area under the precision-recall curve (AUPR) - Top-n accuracy (for multiclass predictions) - Validation on standardized external test set (e.g. ISIC or Melanoma Classification Benchmark)
Model calibration	- Brier score, or mean-squared calibration error - Area under the response rate accuracy curve
Model robustness to "stress tests"	- Stability of prediction as a function of image non-content variance (e.g. rotation, blur, brightness, contrast)

Figure 1. Proposed standardization of metrics for model performance and reporting. Before they can be used effectively in a clinical setting it is proposed that dermatologist-level deep learning models (1) recognize samples that the model is likely to get wrong with the option to abstain from predicting (left panel), (2) report a meaningful confidence associated with each prediction (middle panel), and (3) offer ways to interpret its decision making (right panel). Additionally, it is proposed that models report standard metrics regarding dataset characteristics, model disease discrimination, model calibration, and model robustness to stress tests (box).

and perpetuate inadvertent discriminatory practices, for example, in recommending less follow-up for black patients than for whites, when health costs are used as a proxy for health needs (Obermeyer et al., 2019). Third, disproportionate adoption by different groups may exacerbate existing inequities. Access to and use of technology differs based on sociodemographics (Tsetsi and Rains, 2017), and more tech-savvy users may be more likely to embrace AI for skin screening (Tong and Sopory, 2019). The issue of equity in AI diagnosis needs to be carefully addressed to avoid inadvertent exacerbation of health-care disparities.

Image quality. Several barriers to AI implementation in the clinic need to be overcome with regards to imaging (Figure 1). These include technical variations (e.g., camera hardware and software) and differences in image acquisition and quality (e.g., zoom level, focus, lighting, and presence of hair). For example, the presence of surgical ink markings is associated with decreased specificity (Winkler et al., 2019), field of view can significantly affect prediction quality (Mishra et al., 2019), and classification performance improves when hair and rulers are removed (Bisla et al., 2019). We have developed a method to measure how model predictions might be biased by the presence of a visual artifact (e.g., ink) and proposed methods to reduce such biases (Pfau et al., 2019). Poor quality images are often excluded from studies, but the problem of what makes an image adequate is not well studied. Ideally, models need to be able to express a level of confidence in a prediction as a function of image quality and appropriately direct a user to retake photos if needed.

Model generalizability. Generalizability is a major concern for AI models; studies of computer-assisted diagnosis of melanoma report lower sensitivity for melanoma on independent test sets than on nonindependent test sets (Dick et al., 2019). It is difficult to study generalizability because published DL models are not publicly available, making it impossible to compare performance, unless each study uses a standardized benchmark database, such as the Melanoma Classification Benchmark (Brinker et al., 2019d). Han et al. (2018a) reported excellent metrics of performance and made their model available for image submission; however, the model prediction was not robust when images from an outside clinic were submitted, image magnification or contrast was altered, or images were rotated (Navarrete-Dechent et al., 2018). On ImageNet, a nonmedical dataset of 1,000 object categories, training on a dataset of 300 million unlabeled examples in addition to labeled examples has improved DL model robustness to difficult examples and artificial corruptions (Xie et al., 2019); this method should be tested in dermatology. There is a need for additional standardized benchmark databases spanning different diseases and clinical contexts for use in model performance comparison.

Trusting that an AI model will generalize to a specific patient population ultimately depends on understanding the datasets on which the model was trained and the control experiments that were run (Chuang and Keiser, 2018). For example, otherwise accurate CNNs miss amelanotic melanomas, likely because of underrepresentation in the training set (Tschandl et al., 2019c). Moreover, dermatologists diagnose and manage over 2,000 skin conditions, and although DL algorithms were trained on up to 757 disease classes, they were primarily validated on binary classification tasks (e.g., malignant versus benign); their performance declined markedly when asked to distinguish increased numbers of diagnoses (Esteva et al., 2017). A recent study reported mean top-1 and top-5 model accuracy of 44.8% and 78.1%, respectively, for the classification of 134 diseases (Han et al., 2019b). Most datasets are proprietary, often with minimal description, and datasets collected in dermatology clinics may be skewed toward more complex cases, to those patients with better access to care, or by the choice of camera used in one clinic versus another. Data should be collected from as many diverse sources as possible, including primary care clinics, and robust standards for external validation are needed.

There have been successful efforts to support reproducibility and open access. For example, the study by Han et al. (2018a) details the number and characteristics of images from each data source and makes thumbnails of the images publicly available. Additionally, several studies classifying dermoscopic images use the publicly available International Skin Imaging Collaboration archive (Gutman et al., 2016). By making datasets public, it becomes possible to examine them for bias (Bissoto et al., 2019). Alternatively, reporting a model training database's patient demographics and disease classes would be helpful in predicting model performance on external populations.

Model confidence calibration. It is desirable for an AI model to recognize its limitations and offer a measure of confidence, that is, the probability of being correct, with every prediction. For image classification, models already offer differential diagnoses with varying degrees of confidence rather than making yes or no judgments, but few studies have evaluated how reliably confidence correlates with likelihood of accuracy (Mozafari et al., 2019; Van Molle et al., 2019). Neural network models tend to be overconfident; for example, a model may associate 90% confidence with predictions for which it is correct only 50% of the time (Guo et al., 2017). Thus, before model confidence can be used in practice, it must first be calibrated to accuracy.

Additionally, there is no existing consensus DL technology to understand whether low confidence might reflect an inadequate image (e.g., blurry), an out-of-distribution sample that a model had never encountered during training (e.g., a rare diagnosis or rare visual appearance), or true clinical equivocalness based on visual features. Detecting and distinguishing between these types of difficult images is an active field of DL research. Detecting out-of-distribution samples is particularly challenging. For example, it would be optimal for a model that has only seen pigmented melanomas during training to flag an unfamiliar amelanotic melanoma when it is tasked with making a prediction.

Metrics of model performance. Standard metrics are needed to assess the performance of different models (Figure 1). Currently, standard performance metrics such as accuracy and area under the receiver operating characteristic and precision recall curves are routinely reported. However, for use in the clinic, studies should additionally describe how well their models deal with uncertainty by reporting (i) the Brier Score, or mean-squared calibration error (Rufibach, 2010), which measures how reliably a model can forecast its accuracy, and (ii) area under the response rate accuracy curve, which measures how capably a model can identify examples it is likely to predict falsely and thus abstain from predicting (Hendrycks et al., 2019).

Model interpretability. Acceptance of AI in clinical decision making hinges on being able to understand the decision-making process fundamental to its predictions. DL models are inherently difficult to interpret because they are complex, routinely containing millions of learned parameters; interpretation of DL models' output is an active field of research (Murdoch et al., 2019).

One approach for interpreting model diagnoses is content-based image retrieval, a method for retrieving training images that are visually similar to a test image (Tschandl et al., 2019a). This method may reassure the physician if all the retrieved training images have the same diagnosis as the predicted diagnosis but is less helpful if the test image looks similar to two or more training images with conflicting diagnoses.

A second approach is to highlight pixels in an image most relevant for a model's prediction, using methods such as saliency mapping (Figure 1). However, it is often the case that highlighted pixels correspond to the entire lesion or visually distinctive features that are already obvious to clinicians without indication as to why these pixels are important to the diagnosis.

A third approach is to see through the eyes of a model by plotting an activation atlas (Carter et al., 2019), which shows how subtle changes, in particular visual features, may tip the model over into choosing one diagnosis over another. These activation atlases are experimental and have yet to be applied in dermatology.

Understanding a model's predictions and how the prediction is applicable to the patient at hand is necessary to build trust. As AI exceeds human performance in various tasks, interpreting models may help to advance scientific knowledge by understanding what the machine sees that is relevant to its predications.

CONCLUSION

Automated AI diagnosis of skin lesions is ready to be tested in clinical environments and has potential to provide diagnostic support and expanded access to care. As AI becomes effective at assisting primary care providers with triage through teledermatology, referrals to dermatologists will be for more complex diagnoses and fewer benign diagnoses, such as benign skin lesions. This in turn may cause a contraction in the medical dermatology labor market demand, blunted somewhat by the procedural and cosmetic nature of many practices. We anticipate that dermatologists will see their role shift more toward management of acute and complex skin conditions, including initiating systemic treatment regimens or performing procedures, and involving visits that require face-to-face assessment and/or discussion with regards to patient preferences, values, and logistics. A pilot study indicated that patients preferred to see dermatologists rather than rely only on AI diagnostic support, and patients did not favor replacement of dermatologists by diagnostic support tools (MLW unpublished data). Rather than rendering dermatologists obsolete, in practice, AI may augment dermatologists' clinical assessments in real-time and in teledermatology consults by providing complementary services such as comparing lesions across time and broadening the differential diagnosis. In short, AI may be superb at fast and intuitive pattern recognition but is still far from attaining human-level insight and judgment.

Moreover, there are significant barriers to implementing AI, with technical considerations including model generalizability, confidence calibration, and interpretability. Additional considerations include ensuring equity, defending against security threats, and navigating the regulatory landscape. It is imperative to collect diverse, high-quality datasets for AI training, especially from individuals with darker skin pigmentation underrepresented in current study datasets. Prospective studies are needed to evaluate AI performance (alone and in combination with physicians) compared with standard care. Integrated health-care systems, such as Kaiser Permanente and the Veterans Health Administration, may especially benefit from earlier adoption of AI given cost incentives to reduce unnecessary visits and biopsies of benign lesions.

AI implementation is developing with initial investments from government, industry, and academia, but how AI technologies will be reimbursed is unclear and hinges first on evidence that they improve patient outcomes (He et al., 2019). There is a reimbursement model for teledermatology, using current codes and modifiers, although implementation is patchy, dependent on insurer, and varies by state regulation; modifiers are used to classify visits as live-interactive or store-and-forward teledermatology consults. The reimbursement for the AI portion might take the form of a separate modifier that could be designated for teledermatology or real-time consultation using AI support.

Although AI will be helpful in triaging disease into broad categories with similar treatments, dermatological expertise and clinical correlation will still be needed for fine-grained diagnosis and management decisions or unique cases requiring contextual knowledge (Yu and Wei, 2019). Given the rapid pace of advancements, exposure to the fundamental principles of AI alongside its potential uses and limitations will be crucial for practicing dermatologists and trainees.

ORCIDs

Albert T. Young: http://orcid.org/0000-0002-4088-2488
Mulin Xiong: http://orcid.org/0000-0003-0950-9869
Jacob Pfau: http://orcid.org/0000-0002-5771-8465
Michael J. Keiser: http://orcid.org/0000-0002-1240-2192
Maria L. Wei: http://orcid.org/0000-0002-3568-1921

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

This work was funded by the University of California, San Francisco Helen Diller Family Comprehensive Cancer Center Impact Award (MLW); Melanoma Research Alliance (MLW, ATY, JP); the National Center for Advancing Translational Sciences, National Institutes of Health, through UCSF-CTSI Grant Number 5TL1TR001871-04 (ATY); and the American Skin Association Hambrick Medical Student Grant Targeting Melanoma and Skin Cancer (ATY).

AUTHOR CONTRIBUTIONS

Conceptualization: MLW, ATY; Funding Acquisition: MLW; Project Administration: MLW; Supervision: MLW, MJK; Writing - Original Draft Preparation: ATY, MX; Writing - Review and Editing: ATY, MX, JP, MJK, MLW

REFERENCES

Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol 2018;154:1247–8.

Apple Inc. Core ML. https://developer.apple.com/documentation/coreml; 2020 (accessed 21 November 2019).

Bisla D, Choromanska A, Stein JA, Polsky D, Berman R. Towards automated melanoma detection with deep learning: data purification and augmentation. http://arxiv.org/abs/1902.06061; 2019 (accessed 4 October 2019).

Bissoto A, Fornaciali M, Valle E, Avila S. (De)constructing bias on skin lesion datasets. http://arxiv.org/abs/1904.08818; 2019 (accessed 4 October 2019).

Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer 2019a;119:11–7.

Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019b;113:47–54.

Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer 2019c;111:148–54.

Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer 2019d;111:30–7.

Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. Paper presented at: Conference on Fairness, Accountability and Transparency. 23–24 February 2018; New York, NY.

Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–9.

Carter S, Armstrong Z, Schubert L, Johnson I, Olah C. Activation atlas. Distill 2019;4:e15.

Cho SI, Sun S, Mun JH, Kim C, Kim SY, Cho S, et al. Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network [e-pub ahead of print] Br J Dermatol 2019. https://doi.org/10.1111/bjd.18459 (accessed 9 April 2020).

Chuang KV, Keiser MJ. Adversarial controls for scientific machine learning. ACS Chem Biol 2018;13:2819–21.

Codella NCF, Nguyen Q-B, Pankanti S, Gutman DA, Helba B, Halpern AC, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. http://arxiv.org/abs/1610.04662; 2016 (accessed 4 October 2019).

Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of computer-aided diagnosis of melanoma: a meta-analysis. JAMA Dermatol 2019;155:1291–9.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115−8.

Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24−9.

Ferrante di Ruffano L, Takwoingi Y, Dinnes J, Chuchu N, Bayliss SE, Davenport C, et al. Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults. Cochrane Database Syst Rev 2018;12:CD013186.

Finnane A, Dallest K, Janda M, Soyer HP. Teledermatology for the diagnosis and management of skin cancer: a systematic review. JAMA Dermatol 2017;153:319−27.

Freeman K, Dinnes J, Chuchu N, Takwoingi Y, Bayliss SE, Matin RN, et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. BMJ 2020;368:m127.

Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol 2019;180:373−81.

Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. http://arxiv.org/abs/1706.04599; 2017 (accessed 4 October 2019).

Gutman D, Codella NCF, Celebi E, Helba B, Marchetti M, Mishra N, et al. Skin lesion analysis toward melanoma detection: a challenge at the International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). http://arxiv.org/abs/1605.01397; 2016 (accessed 4 October 2019).

Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29:1836−42.

Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. Ann Oncol 2020;31:137−43.

Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018a;138:1529−38.

Han SS, Moon IJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network [epub ahead of print] JAMA Dermatol 2019. https://doi.org/10.1001/jamadermatol.2019.3807 (accessed 9 April 2020).

Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PLoS One 2018b;13:e0191493.

Han SS, Park I, Chang S, Na J. Deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for general skin disorders. J Invest Dermatol 2019b;139(Suppl):S171.

He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30−6.

Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer 2019a;120:114−21.

Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019b;118:91−6.

Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D. Natural adversarial examples. http://arxiv.org/abs/1907.07174; 2019 (accessed 21 November 2019).

Ianni JD, Soans RE, Sankarapandian S, Chamarthi RV, Ayyagari D, Olsen TG, et al. Augmenting the pathology lab: an intelligent whole slide image classification system for the real world. http://arxiv.org/abs/1909.11212; 2019 (accessed 15 October 2019).

Jayakumar KL, Lipoff JB. Trends in the dermatology residency match from 2007 to 2018: implications for the dermatology workforce. J Am Acad Dermatol 2019;80:788−90.

Khullar DAI. could worsen health disparities. https://www.nytimes.com/2019/01/31/opinion/ai-bias-healthcare.html; 2019. (accessed October 4, 2019).

Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Heal 2019a;1:e271−97.

Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. http://arxiv.org/abs/1909.05382; 2019b (accessed 4 October 2019).

Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78:270−7.e1.

Marchetti MA, Liopyris K, Dusza SW, Codella NCF, Gutman DA, Helba B, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the international skin imaging collaboration 2017. J Am Acad Dermatol 2020;82:622−7.

Maron RC, Weichenthal M, Utikal JS, Hekler A, Berking C, Hauschild A, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Cancer 2019;119:57−65.

Martinka MJ, Crawford RI, Humphrey S. Clinical recognition of melanoma in dermatologists and nondermatologists. J Cutan Med Surg 2016;20:532−5.

Mishra S, Imaizumi H, Yamasaki T. Interpreting fine-grained dermatological classification by deep learning. CVPR Workshops 2019. https://www.semanticscholar.org/paper/Interpreting-Fine-Grained-Dermatological-by-Deep-Mishra-Imaizumi/72d0b1f44a9bcc5fa2588df36c8e568190dc79d6 (accessed 4 October 2019).

Mozafari AS, Gomes HS, Leão W, Gagné C. Unsupervised temperature scaling: an unsupervised post-processing calibration method of deep networks. https://arxiv.org/abs/1905.00174v3; 2019 (accessed 22 January 2020).

Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci USA 2019;116:22071−80.

National Cancer Institute. Cancer stat facts: melanoma of the skin. https://seer.cancer.gov/statfacts/html/melan.html; 2020 (accessed October 4, 2019).

Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality? J Invest Dermatol 2018;138:2277−9.

Navarro F, Escudero-Vinolo M, Bescos J. Accurate segmentation and registration of skin lesion images to evaluate lesion change. IEEE J Biomed Health Inform 2019;23:501−8.

Naylor CD. On the prospects for a (deep) learning health care system. JAMA 2018;320:1099−100.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447−53.

Onega T, Barnhill RL, Piepkorn MW, Longton GM, Elder DE, Weinstock MA, et al. Accuracy of digital pathologic analysis vs traditional microscopy in the interpretation of melanocytic lesions. JAMA Dermatol 2018;154:1159−66.

Pfau J, Young AT, Wei ML, Keiser MJ. Global saliency: aggregating saliency maps to assess dataset artefact bias. https://arxiv.org/abs/1910.07604; 2019 (accessed 17 October 2019).

Phillips M, Marsden H, Jaffe W, Matin RN, Wali GN, Greenhalgh J, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. JAMA Netw Open 2019;2:e1913436.

Piepkorn MW, Longton GM, Reisch LM, Elder DE, Pepe MS, Kerr KF, et al. Assessment of second-opinion strategies for diagnoses of cutaneous melanocytic lesions. JAMA Netw Open 2019;2:e1912597.

Rufibach K. Use of Brier score to assess binary predictions. J Clin Epidemiol 2010;63:938−9 [author reply 939].

Szyc Ł, Hillen U, Scharlach C, Kauer F, Garbe C. Diagnostic performance of a support vector machine for dermatofluoroscopic melanoma recognition: the results of the retrospective clinical study on 214 pigmented skin lesions. Diagnostics (Basel) 2019;9:103.

TensorFlow. https://www.tensorflow.org/lite; 2020 (accessed 21 November 2019).

Tong ST, Sopory P. Does integral affect influence intentions to use artificial intelligence for skin cancer screening? A test of the affect heuristic. Psychol Health 2019;34:828–49.

Tschandl P, Argenziano G, Razmara M, Yap J. Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features. Br J Dermatol 2019a;181:155–65.

Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol 2019b;20:938–47.

Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol 2019c;155:58–65.

Tsetsi E, Rains SA. Smartphone Internet access and use: extending the digital divide and usage gap. Mob Media Commun 2017;5: 205015791770832.

Van Molle P, Verbelen T, De Boom C, Vankeirsbilck B, De Vylder J, Diricx B, et al. Quantifying uncertainty of deep neural networks in skin lesion classification. In: Greenspan H, Tanno R, Erdt M, Arbel T, Baumgartner C, Dalca A, et al., editors. Uncertainty for safe utilization of machine learning in medical imaging and clinical image-based procedures. Cham, Switzerland: Springer International Switzerland AG; 2019. p. 52–61.

VisualDx. Dermatology. https://www.visualdx.com/professionals/dermatology; 2020 (accessed 4 Oct 2019).

Wilmer EN, Gustafson CJ, Ahn CS, Davis SA, Feldman SR, Huang WW. Most common dermatologic conditions encountered by dermatologists and nondermatologists. Cutis 2014;94:285–92.

Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol 2019;155:1135–41.

Xie Q, Hovy E, Luong M-T, Le QV. Self-training with noisy student improves ImageNet classification. http://arxiv.org/abs/1911.04252; 2019 (accessed 11 November 2019).

Xiong M, Pfau J, Young AT, Wei ML. Artificial intelligence in teledermatology. Curr Dermatol Rep 2019;8:85–90.

Yu WY, Wei ML. Suction blisters. JAMA Dermatol 2019;155:237.