Deep Learning in Dermatology: A Systematic Review of Current Approaches, Outcomes, and Limitations

Hyeon Ki Jeong[1,2], Christine Park[1], Ricardo Henao[2] and Meenal Kheterpal[1]

Artificial intelligence (AI) has recently made great advances in image classification and malignancy prediction in the field of dermatology. However, understanding the applicability of AI in clinical dermatology practice remains challenging owing to the variability of models, image data, database characteristics, and variable outcome metrics. This systematic review aims to provide a comprehensive overview of dermatology literature using convolutional neural networks. Furthermore, the review summarizes the current landscape of image datasets, transfer learning approaches, challenges, and limitations within current AI literature and current regulatory pathways for approval of models as clinical decision support tools.

JID Innovations (2023);**3:**100150 doi:10.1016/j.xjidi.2022.100150

INTRODUCTION

Artificial intelligence (AI) in healthcare is the application of machine learning (ML) algorithms in medical fields to potentially improve diagnosis and predict clinical outcomes (Jiang et al., 2017). The advancements in computing power and vast data curation within health systems have led to algorithm development that can assist healthcare providers as clinical decision-support (CDS) tools. A myriad of AI applications exists within health care such as using electronic health record data for risk predictors (Juhn and Liu, 2020; Lauritsen et al., 2020), early prediction and diagnosis of diseases such as sepsis (Goh et al., 2021; Komorowski et al., 2018), and continuous disease monitoring using wearable devices. There have been innovative efforts to procure large numbers of medical image datasets, either within institutions or for public use, such as DeepLesion, which contains 32,000 computed tomography images for scientific studies (Yan et al., 2018), or the National Institutes of Health Chest X-Ray Dataset (Wang et al., 2017[1]).

Computer vision is a field of AI in which the system learns to interpret visual images. It has advanced the process of medical image evaluation with higher accuracy and more efficient analysis (Voulodimos et al., 2018). The convolutional neural network (CNN) is a type of artificial neural network that has revolutionized image analysis without the need to extract traditional handcrafted features such as colors, intensity value, topological structure, and texture information (Carin and Pencina, 2018). Researchers have developed deep learning models that have been trained on millions of images for different tasks such as image classification, object detection, and image recognition. Model development for computer vision challenges such as image classification and objection detection is achieved by training and testing on millions of images. These models, most notably inspired by ImageNet (Deng et al., 2009), CIFAR (Krizhevsky and Hinton, unpublished data), Modified National Institute of Standards and Technology (MNIST) (Deng, 2012), COCO (Common Objects in Context) (Lin et al., 2014), Open Images (Kuznetsova et al., 2020[2]), and SUN (Xiao et al., 2010) challenges, can either detect or classify numerous different categories such as dogs or cats in a given image with high accuracy.

Medical imaging field has adapted these CNN methods to solve a diverse array of problems, using datasets obtained from various imaging modalities such as chest x-rays (Lakhani and Sundaram, 2017), magnetic resonance imaging (Pereira et al., 2016), pathology (Kermany et al., 2018), and ophthalmology (Gulshan et al., 2016). In medical image analysis, the lack of data creates a bottleneck for training a deep learning model. Acquiring and annotating medical images is costly, time consuming, and labor intensive. Data sharing may serve as a potential solution to accelerate data collection, but ethical and privacy issues can hinder institutional data sharing. Hence, transfer learning has vastly improved the medical imaging field by allowing the use of models that have been pretrained on millions of images to solve numerous medical imaging problems, alleviating the need to spend hours building an effective model or collecting vast amounts of clinical data.

[1]*Duke Dermatology, Duke University School of Medicine, Durham, North Carolina, USA; and* [2]*Department of Biostatistics & Bioinformatics, Duke University School of Medicine, Durham, North Carolina, USA*

Correspondence: Meenal Kheterpal, Duke Dermatology, Duke University School of Medicine, 40 Duke Medicine Circle, Durham, North Carolina 27710, USA. E-mail: meenal.kheterpal@duke.edu

Abbreviations: AI, artificial intelligence; CDS, clinical decision-support; CNN, convolutional neural network; FDA, Food and Drug Administration; ISIC, International Skin Imaging Collaboration; ML, machine learning; MNIST, Modified National Institute of Standards and Technology; SaMD, Software as a Medical Device; SVM, support vector machine

Received 30 April 2022; revised 17 June 2022; accepted 15 July 2022; corrected proof published online XXX

Cite this article as: JID Innovations 2023;3:100150

[1] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv 2017.

[2] Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, et al. The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. arXiv 2020.

Pretrained models can be fine tuned to unique problems according to the amount of available data and the data similarity.

Thus, one can choose a standard model, often trained on a popular dataset (such as ImageNet), and fine tune the network to fit a given problem. Contrary to the assumptions that weights from a model pretrained on real-world images may not translate well for medical images, studies have shown that ImageNet-pretrained models have produced human-level accuracy for medical image classification such as in pathology (Ehteshami Bejnordi et al., 2016; Gown et al., 2008; Qaiser and Mukherjee, 2018) and dermatology (Cho et al., 2020; Haenssle et al., 2018; Maron et al., 2019). In dermatology, AI systems using transfer learning are comparable with or even surpass the performance of dermatologists in diagnosing skin conditions (Esteva et al., 2017; Haenssle et al., 2020). The rapidly growing global burden of skin cancer, rise of teledermatology during the COVID-19 pandemic, and supply–demand imbalance for dermatologists point to an escalating need to establish effective triaging systems supported by AI for dermatological disease detection and diagnosis. This study aims to provide a comprehensive review of published applications of pretrained models on dermatological images, their associated datasets, their limitations, and their outcomes.

Search strategy

We conducted a query of MEDLINE and PubMed Central databases through PubMed using keywords, including dermatology, deep learning, transfer learning, and convolutional neural network. Studies published from the year 2016 to 2021 were included. We excluded studies on the basis of the following criteria: (i) use of non–deep learning algorithms, (ii) use of custom algorithms, and (iii) meta-analysis and/or review articles. The search strategies are outlined in Figure 1.

RESULTS

In total, 65 studies were included in the final review. Table 1 summarizes the classification tasks, methodology, and outcome metrics for the CNNs developed for skin conditions. Table 2 provides details of the publicly available dataset investigated in this review. Table 3 classifies the studies according to the type of dataset (i.e., institutional or public) and image (i.e., clinical or dermoscopic).

We found 22 different types of the pretrained models used for dermatology application, with ResNet being the most widely used, followed by Inception and VGG. A total of 45 studies conducted classification tasks across different skin diseases, including melanoma, foot ulcer, psoriasis, and rosacea. A total of nine studies focused on skin lesion segmentation and two studies focused on skin lesion detection (bounding box) mostly using U-Net architecture, which is well-suited for object detection tasks. A total of six studies tackled both segmentation and classification tasks separately, whereas three studies aimed to develop an end-to-end model from segmentation to classification.

DISCUSSION

Model selection and feature extraction approaches

The classification methods can be divided into two approaches: single deep learning models and ensemble methods (Dietterich, 2000).

Single deep learning models, as the name implies, use a single pretrained model without modification of the architecture. Often, the studies tested multiple models and report on the one with the best performance. For instance, Yap et al. (2018) investigated five different models, such as VGG16, ResNet-101, InceptionV3, DenseNet121, and EfficientNet, for classifying infection and ischemia of diabetic foot ulcers and reported that the EfficientNetB0 had the best results. Guergueb and Akhloufi (2021) evaluated various submodels of VGG, ResNet, EfficientNet, DenseNet, Inception, and MobileNet for binary melanoma classification and discovered that EfficientNetB7 had the highest accuracy of 99.33%.

The ensemble method combines predictions from two or more models that could improve the predictive performance instead of a single model (Sagi and Rokach, 2018). Harangi (2017)[3] used an ensemble of GoogLeNet, AlexNet, ResNet-50, and VGG-VG-16 for melanoma classification and showed that the ensemble method outperforms each individual deep learning method. Han et al. (2018) used an ensemble of ResNet-152 and VGG19 for onychomycosis diagnosis and achieved the highest classification performance than the dermatologists.

Few studies used pretrained models to extract features and apply other traditional classification algorithms such as support vector machine (SVM) or XGBoost. Mahbod et. al. (2019) used AlexNet, VGG16, and ResNet-18 as feature extractors and used an SVM as a classifier for each network. Each SVM score is fused to obtain a probability for binary classification. Yu et. al. (2017) used ResNet-50 to extract features and averaged the scores from a neural network classification layer and the SVM classifier to obtain the final prediction. Tschandl et al. (2019) combined outputs from InceptionV3 and ResNet-50 and used XGBoost to compute the probabilities.

Datasets in dermatology

The key to developing a high-performance deep learning model is the data. If the number of high-quality datasets is large, there is a higher likelihood that the models will learn to generate more accurate predictions. Several publicly available skin image datasets are provided to engage both dermatology and ML communities to develop novel or to hone existing algorithms. For example, the International Skin Imaging Collaboration (ISIC) archive is one of the most well-known public skin cancer image datasets that has gained a high reputation over the years through algorithmic challenges such as lesion segmentation, visual dermoscopic feature detection and localization, and disease classification since 2016 (Codella et al., 2018, 2017[4]; Tschandl et al., 2018). The archive contains over 13,000 dermoscopic images collected

[3] Harangi B. Skin lesion detection based on an ensemble of deep convolutional neural network. arXiv 2017.
[4] Codella NCF, Nguyen Q-B, Pankanti S, Gutman DA, Helba B, Halpern AC, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. arXiv 2017.

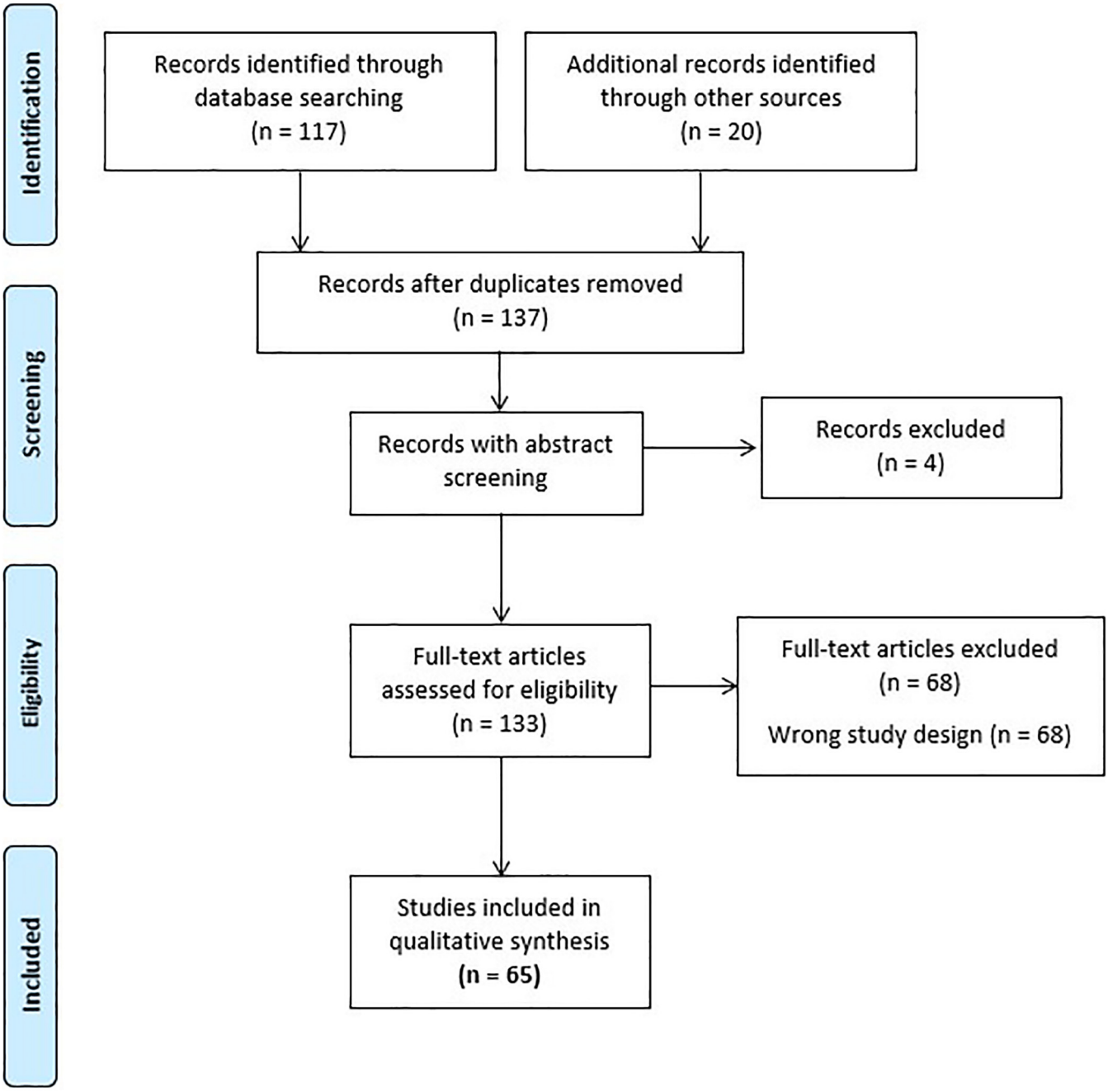

Figure 1. PRISMA diagram

from leading clinical centers internationally. Additional image repositories include dermatology atlases that were originally used for educational purposes but have recently been used as a database of digital images for algorithm development (Tables 2 and 3). There are few datasets that are available upon a fee and a licensing agreement such as Dermofit Image Library or an ethical committee/institutional approval (Han et al., 2020a; Papadakis et al., 2021; Webster et al., 2017). Other public datasets of different skin diseases include the diabetic foot ulcer challenge, providing >15,000 images of diabetic foot ulcers, other foot/skin conditions, and healthy feet taken with three digital cameras. Besides these public datasets, many clinical institutions have collected their own respective datasets for diseases such as psoriasis, rosacea, and lip disorder.

Challenges with image data/datasets

Duplicity of data. Cassidy et al. (2022) noted that several manuscripts using the ISIC dataset had duplicate or similar images within training and test sets, introducing bias into the CNN model, and proposed a methodology to remove duplicate images. Because the model predictions improve in accuracy by extracting a higher number of unique features rather than by simply enriching the data by sourcing a large number of images from a small number of sources, it must be noted that a rich source of nonduplicated data must be used

Table 1. Summary of Tasks, Methodology, Dataset, and Performance

Author	Objective	Model Tested	Dataset	Model Performance	Limitations/Comments
Melanoma					
Bi et al., 2017[7]	Segmentation/classification (separate) Multiclass: melanoma, seborrheic keratosis, and nevus	ResNet	Public/dermoscopy ISIC 2017 skin lesion analysis challenge	Segmentation Jaccard Index 79.4 Classification AUC: 0.843 SE: 69.3% SP: 83.6%	N/A
Shahin et al., 2018	Classification Binary: malignant (melanoma) versus nonmalignant (nevi)	Ensemble of ResNet-50 and Inception V3	Public/dermoscopy ISIC 2018: Skin lesions analysis toward melanoma detection	Acc: 89.9% Average precision of 0.862 Average recall of 0.796	Data imbalance (used weighted cross entropy to alleviate effects) Did not manage to run cross validation to ensure stability and robustness
Yap et al., 2018	Classification Binary: Melanoma, seborrheic keratosis, and nevus	Two ResNet50 fusion	Institutional/dermoscopy, clinical, meta-data Multiple skin cancer clinics (macroscopic image, dermatoscopic image, meta-data) Image label split: 350 Acral melanoma; 374 acral nevi	Best results AUC of dsc + macro: 0.866 mAP of dsc + macro + meta: 0.729	Lack of patient information/ clinical information Adding age, gender, location, and lesion size increased the accuracy Common verification bias in dermatoscopic studies, with only pathologically diagnosed cases included.
Brinker et al., 2019a	Classification Binary: melanoma versus atypical nevi	ResNet50	Public/clinical Trained with ISIC image archive and HAM1000 dataset Validated with Mclass-Benchmark for clinical images obtained from MED-NODE database	SE: 89.4% SP: 68.2%	N/A
Brinker et al., 2019b	Classification Binary: malignant (melanoma) versus nonmalignant (nevi)	ResNet50	Public/dermoscopy ISIC image archive 2018	At SE: 74.1% model achieved SP: 86.5% At SP: 69.2% model achieved SE: 84.5%	Clinical encounters with actual patients provide more information than can be provided by images alone
Hekler et al., 2019	Classification Binary: melanoma versus nevi	ResNet50	Public/dermoscopy Trained with ISIC image archive and HAM1000 dataset Tested on biopsy-verified images from HAM1000 dataset	Acc: 81.59% SE: 86.1% SP: 89.2%	Lack of patient information for AI models' algorithms performance would be worse on an entirely external dataset of images.
Salamaa and Aly, 2021	Classification Binary: malignant (melanoma) versus nonmalignant (nevi)	Tested VGG16 and ResNet50 with SVM ResNet50 showed the best performance	Public/dermoscopy ISIC2017, MNIST-HAM10000, and ISBI 2016	Acc: 99.19% AUC: 99.32% SE: 98.98% Precision: 98.78% F1 score: 98.88%	N/A
Jojoa Acosta et al., 2021	Classification Binary: benign versus malignant (melanoma, seborrheic keratosis, and nevus)	ResNet152	Public/dermoscopy ISIC challenge 2017	SE: 0.820 SP: 0.925 Acc: 0.904 Balanced Acc: 0.872	N/A
Yu et al., 2017	Segmentation/classification (separate) Binary: malignant versus nonmalignant	Segmentation: fully convolutional residual network Classification: very deep CNNs (residual) with softmax and SVM classifier	Public/dermoscopy ISBI 2016 Skin lesion analysis toward melanoma detection challenge Image data split: not reported	Segmentation: Acc: 0.949 SE: 0.911 SP: 0.957 Classification: Acc 0.855 SE: 0.547 SP: 0.931	Insufficiency of quality training data Difficulty in fully exploiting the discrimination capability gains of very deep CNNs under the circumstance of limited training data
Harangi, 2017	Segmentation/classification (separate) Multiclass: melanoma, seborrheic keratosis, and nevus	Ensemble of GoogLeNet, AlexNet, ResNet50, and VGG-VD-16	Public/dermoscopy ISIC 2017 skin lesion analysis challenge	AUC: 0.932 SE: 82% | SP: 89.4% SE: 89% | SP: 85.0% SE: 95% | SP: 65.9%	N/A
Li and Li, 2018[8]	Segmentation/classification (end-to-End) Multiclass: melanoma, AK, nevus, BCC, dermatofibroma, vascular lesion, and benign keratosis	Segmentation: ResNet Classification: ResNet, DenseNet, Inception	Public/dermoscopy ISIC challenge 2018 HAM10000 dataset	Segmentation Jaccard Index: 0.818 Classification DenseNet121: 0.848 ResNet152: 0.86 Inception, version 4: 0.85	N/A

(*continued*)

Table 1. Continued

Author	Objective	Model Tested	Dataset	Model Performance	Limitations/Comments
				DenseNet w/cropped image: 0.912	
Rezvantalab et al., 2018[9]	Classification Binary: melanoma, melanocytic nevi, BCC, benign keratosis, AK and intraepithelial carcinoma, dermatofibroma, vascular lesions, and atypical nevi	Comparing DenseNet 201; ResNet152; Inception, version 3; and InceptionResNet, version 2	Public/dermoscopy HAM10000 Dataset PH2 dataset	AUC of melanoma versus that of basal cell carinoma 94.40% (ResNet152) 99.3% (DensNet 201)	The utilized dataset is highly unbalanced, and also no preprocessing step is applied in this paper. Still the results are promising
Mahbod et al., 2019	Classification Binary: melanoma and seborrheic keratosis versus nevus	AlexNet, VGG16, and ResNet18 for feature extraction SVM for classification	Public/dermoscopy ISIC 2016 and 2017 competition	AUC: 83.83% (melanoma) AUC: 97.55% (seborrheic keratosis)	N/A
Tschandl et al., 2019	Classification Binary: malignant versus benign nonpigmented skin lesions	Combined Inception, version 3 (dermoscopic images), and ResNet50 (clinical close-ups) using xgboost	Institutional/dermoscopy and clinical 7,895 dermoscopic and 5,829 close-up images of the training set originated from a consecutive sample of lesions photographed and excised by one author (CR) at a primary skin cancer clinic in Queensland, Australia. Image data split: described in the manuscript in detail	AUC: 0.742 SE: 80.5% SP: 53.5%	Test set included >51 distinct classes, of which most did not have enough examples to be integrated into the training phase
Chang, 2017[10]	Segmentation/classification (separate) Binary: malignant versus benign nonpigmented skin lesions	Segmentation: U-Net Classification: Google Inception, version 3	Public/dermoscopy ISIC challenge website. A total of 2,000 dermoscopic images includes 374 melanoma images, 1,372 nevus images, and 254 seborrheic keratosis images	N/A	N/A
Esteva et al., 2017	Classification keratinocyte carcinomas versus benign seborrheic keratoses and malignant melanomas versus benign nevi.	Inception, version 3, CNN architecture	Institutional and public/dermoscopy ISIC dermoscopic archive, the Edinburgh Dermofit Library 22, and data from the Stanford Hospital	AUC for different image sets carcinoma 135 images: 0.96 carcinoma 707 images: 0.96 melanoma 130 images: 0.94 melanoma 225 images: 0.94 melanoma 111 dermoscopy images: 0.91 melanoma 1,010 dermoscopy images: 0.94	The CNN achieves performance on par with all tested experts across both tasks, showing an AI capable of classifying skin cancer with a level of competence comparable with that of dermatologists
Mirunalini et al., 2017[11]	Classification Binary: melanoma, seborrheic keratosis, and nevus	Google Inception, version 3	Public/dermoscopy ISIC challenge 2017	Acc: 65.8%	N/A
Murphree and Ngufor, 2017[12]	Classification Binary: melanoma, seborrheic keratosis, and nevus	Google Inception, version 3	Public/dermoscopy ISIC challenge 2017	AUC: 0.84 (nevus and seborrheic keratosis) AUCL 0.76 (melanoma)	N/A
Haenssle et al., 2018	Classification Binary: melanoma versus melanocytic nevi	Google's Inception, version 4	Public and Institutional/dermoscopy Trained and validated on ISIC and ISBI 2016 dataset Tested on 300 from the image library of the Department of Dermatology, University of Heidelberg (Heidelberg, Germany) and on 100 images from ISIC and ISBI dataset Image data split: 20% melanoma, 80% benign nevi	300 test set SE: 95% SP: 80% AUC: 0.95 100 Test Set SE: 95% SP: 63.8% AUC: 0.86	Lack of melanocytic lesions from other skin types
Kawahara et al., 2018	Classification Multiclass: melanoma, melanocytic nevi, BCC, benign keratosis, AK and intraepithelial carcinoma, dermatofibroma, vascular lesions, and atypical nevi	Inception v3	Public/dermoscopy, clinical, meta-data Interactive Atlas of Dermoscopy by Argenziano (made publicly available)	SE: 60.4 SP: 91.0 Precision: 69.6 AUC: 89.6	N/A

(*continued*)

Table 1. Continued

Author	Objective	Model Tested	Dataset	Model Performance	Limitations/Comments
Winkler et al., 2019	Classification Binary: malignant (melanoma) versus nonmalignant (nevi)	Inception, version 4	Institutional/dermoscopy Department of Dermatology, University of Heidelberg	Unmarked lesion SE: 95.7% SP: 84.1% AUC: 0.969 Marked lesion SE: 100% SP: 45.8% AUC: 0.922 Cropped image SE: 100% SP: 97.2% AUC: 0.993	Most images included in this study were derived from fair-skinned patients residing in Germany; therefore, the findings may not be generalized for lesions of patients with other skin types and genetic backgrounds.
Fujisawa et al., 2019	Classification Binary: benign versus malignant lesions	GoogLeNet	Institutional/dermoscopy 4,867 clinical images obtained from 1,842 patients diagnosed with skin tumors from the University of Tsukuba Hospital	Acc: 76.5% SE: 96.3% SP: 89.5%	N/A
Vasconcelos and Vasconcelos, 2017[13]	Classification Binary: melanoma, seborrheic keratosis, and nevus	GoogLeNet	Public/dermoscopy ISIC challenge 2017	AUC: 0.932	N/A
Sousa and de Moraes, 2017[14]	Classification Binary: melanoma, seborrheic keratosis, and nevus	GoogLeNet AlexNet	Public/dermoscopy ISIC challenge 2017	AUC: 0.95 (GoogLeNet) AUC: 0.846 (AlexNet)	N/A
Yang et al., 2017[15]	Segmentation/classification (separate) Binary: melanoma, seborrheic keratosis, and nevus	Segmentation: U-Net Classification: GoogLeNet	Public/dermoscopy ISIC challenge 2017	Segmentation: Jaccard Index: 0.724 Classification: AUC: 0.880 0.972	N/A
Codella et al., 2017	Segmentation/classification (end-to-End) Binary: melanoma versus melanocytic nevi	Segmentation: similar to U-Net architecture Classification: ensemble of deep residual network, CaffeNet, U-Net architecture	Public/dermoscopy 900 training and 379 testing images from ISBI 2016 dataset	Segmentation Jaccard Index 0.84 Acc: 95.1% Classification AUC: 0.843 SE: 69.3% SP: 83.6%	Lack of patient information for AI models
Ashraf et al., 2022	Segmentation Automated prediction of lesion segmentation boundaries from dermoscopic images	UNet, deep residual U-Net (ResUNet), and improved ResUNet (ResUNet++)	Public/dermoscopy ISIC 2016 and 2017 database	Jaccard Index 80.73% on ISIC 2016 90.02% on ISIC 2017	N/A
Mishra and Daescu, 2017	Segmentation Automated prediction of lesion segmentation boundaries from dermoscopic images	U-Net	Public/dermoscopy ISIC 2017	Jaccard Index: 0.842 Acc: 0.928 SE: 0.930 SP: 0.954 Dice Coeff: 0.868	N/A
Araújo et al., 2021	Segmentation Automated prediction of lesion segmentation boundaries from dermoscopic images	U-Net	Public/dermoscopy PH2 and DermIS	Dice Coeff PH2: 0.933 DermIS: 0.872	N/A
Pomponiu et al., 2016	Classification Binary: malignant versus nonmalignant	AlexNet	Public/dermoscopy DermIS and DermQues	SE: 92.1 SP: 95.18 Acc: 93.64	N/A
Pour et al., 2017	Segmentation Automated prediction of lesion segmentation boundaries from dermoscopic images	FCN-AlexNet with seven convolutional layers and a deeper model that is VOC-FCN8s with 15 convolutional layers	Public/dermoscopy ISBI 2016 Skin lesion analysis toward melanoma detection challenge	SE: 0.91 SP: 0.95 Acc: 0.94 JA: 0.83 DI: 0.89	N/A
Kaymak et al., 2018	Classification Binary: malignant (melanoma) versus nonmalignant (nevi)	AlexNet	Public/dermoscopy ISIC 2018. Skin lesions analysis toward melanoma detection	Acc: 84% SE: 84.7% SP: 83.8%	N/A

(*continued*)

Table 1. Continued

Author	Objective	Model Tested	Dataset	Model Performance	Limitations/Comments
Menegola et al., 2017	Classification Melanoma, BCC, nevus	VGG-M with SVM as a classifier	Public/dermoscopy Interactive atlas of dermoscopy (atlas), and the ISBI challenge 2016/ISIC	AUC: 80.7% SE: 47.6% SP: 88.1%	N/A
Yu et al., 2018	Classification Binary: acral melanoma and benign nevi	Fine-tuned modified VGG model with 16 layers	Institutional/dermoscopy A total of 724 dermoscopy images were collected from January 2013 to March 2014 at the Severance Hospital in the Yonsei University Health System (Seoul, Korea) and from March 2015 to April 2016 at the Dongsan Hospital in the Keimyung University Health System (Daegu, Korea)	Acc: Group A: 83.51% Group B: 80.23% AUC Group A: 0.8 Group B: 0.84	N/A
Lopez et al., 2017	Classification Binary: malignant versus nonmalignant	VGGNet	Public/dermoscopy ISBI 2016 Skin lesion analysis toward melanoma detection challenge	SE: 78.66% Precision: 0.7974 Acc: 81.33%	N/A
Shorfuzzaman, 2022	Classification Binary: malignant (melanoma) versus nonmalignant (nevi)	Ensemble of EffcientNetB0, DenseNet121, and Xception	Public/dermoscopy PH2 and ISIC (018 and 2019 database	Acc: 95.76% SE: 96.67% AUC: 0.957	N/A
Guergueb and Akhloufi, 2021	Classification Binary: malignant (melanoma) versus nonmalignant (nevi)	Multiple models were tested, and Efficientnet b7 showed the best result	Public/dermoscopy SIIM-ISIC 2020, ISIC's archive, ISIC 2019, ISIC 2018, and ISIC 2017	Acc: 99.33 SE: 98.78 SP: 99.38 AUC: 99.01	N/A
Han et al., 2020b	Detection/classification (end-to-end) Binary: Benign versus malignant lesions	Detection: Faster R-CNN Classification: SENet, SE-ResNeXt-50, and SE-ResNet-50	Institutional/clinical 1,106,886 training set, 2,844 validation set, and 325 test set from Asan Medical Center; plastic surgery from Chonnam National University Department of Plastic Surgery and Hallym University Department of Plastic Surgery	AUC: 0.92 SE: 92.5% at t > 0.9 SP: 70.0% at t > 0.9 SE: 80.0% SS at t > 0.8 SP: 87.5% at t > 0.8	Algorithm was validated with one race (Asian) within one region (South Korea). Model only has photo information and lacks other evaluations from physicians.
Li and Shen, 2018	Segmentation/Classification (separate) Binary: melanoma versus seborrheic keratosis and nevus	FCRN-88	Public/dermoscopy ISIC 2017 skin lesion analysis challenge	Segmentation Jaccard Index 0.753 Classification AUC: 0.912	N/A
Jafari et al., 2017	Segmentation Automated prediction of lesion segmentation boundaries from dermoscopic images	Basic architecture of the configured CNN is inspired by the layers in the LeNet network	Public/dermoscopy Dataset of skin lesion images from Dermquest database that is publicly available with segmentation ground truth	Melanoma lesion SE: 95.2% SP: 99.0% Acc: 98.7%	N/A
Goyal et al., 2019	Segmentation Automated prediction of lesion segmentation boundaries from dermoscopic images	ResNet-Inception, version 2; DeepLab, version 3 Ensemble-A: combines results from both models Ensemble-L: chooses the larger area Ensemble-S: chooses a smaller area	Public/dermoscopy PH2 and ISIC 2017	Ensemble-A (best performance) Acc: 0.941 Dice: 0.871 Jaccard Index: 0.793 SE: 0.899 SP: 0.950	N/A
Psoriasis					
Zhao et al., 2020	Classification Design and evaluation of a smart psoriasis identification system based on clinical images	DenseNet; Inception, version 3; InceptionResNet, version 2; and Xception Inception, version 3, performed best	Institutional/clinical Images collected by dermatologists at Xiangya Hospital, Annotated by three dermatologists with >10 years' experience at Xiangya Hospital according to the corresponding	AUC: 0.981 ± 0.015 SE: 0.98 SP: 0.92	Model has the capability to identify psoriasis (acc of model: 0.96) with a level of competence comparable with those of 25 dermatologists (mean acc of 25 dermatologists: 0.87)

(*continued*)

Table 1. Continued

Author	Objective	Model Tested	Dataset	Model Performance	Limitations/Comments
			medical record and pathology results		
Dash et al., 2019	Segmentation CNN model for detection of psoriasis lesions	U-Net	Institutional/dermoscopy Images captured and annotated by a dermatologist at the Psoriasis Clinic and Research Centre, Psoriatreat, Pune, Maharashtra, India	Acc: 94.80 Dice Coeff: 93.03 Jaccard Index: 86.40 SE: 89.60 SP: 97.60	N/A
Raj et al., 2021	Segmentation Automatic approach based on a deep learning model using transfer learning for the segmentation of psoriasis lesions from the digital images of different body regions of patients with psoriasis.	U-Net	Institutional/clinical Psoriasis Clinic and Research Centre, Psoriatreat, Pune, Maharashtra, India	Dice similarity: 0.948 Jaccard Index: 0.901	N/A
Yang et al., 2021	Classification Train an efficient deep-learning network to recognize dermoscopic images of psoriasis (and other papulosquamous diseases), improving the Acc of the diagnosis of psoriasis	EfficientNet-b4	Institutional/dermoscopy 7,033 dermoscopic images from 1,166 patients collected from the Department of Dermatology, Peking Union Medical College Hospital (Peking, China)	Psoriasis SE: 0.929 SP: 0.952 Eczema SE: 0.773 SP: 0.926 Lichen planus SE: 0.933 SP: 0.960 Other groups SE: 0.840 SP: 0.985	The algorithm only recognized the dermoscopic images from lesions to diagnose the disease, different from the clinical diagnosis process with multimodal data (e.g., age, sex, medical history, and treatment response) and more types of diseases involved. The model may not perform well on other populations with different skin types/colors.
Meienberger et al., 2020	Classification To establish psoriasis assessment on the basis of segmenting images using machine learning	A fully convolutional neural network called Net16 uses a residual connection architecture as introduced by He et al. (2016)	Institutional/clinical 203 photographs of Caucasian patients aged between 18 and 80 years and suffering from plaque-type psoriasis were selected. The photographs included were taken with a Nikon D700 camera	Acc: 0.91 F1-score: 0.71	Restriction of the data is the inclusion of mostly Caucasian patients. Because the manifestation of psoriasis differs depending on the skin type, including only a few images of other skin types would have led to a highly imbalanced data set
Arunkumar and Jayanna, 2021	Classification Automatically classify psoriasis-affected skin area from normal healthy skin using machine learning algorithm	mobilenet, nasnetlarge NasNetLarge chosen	Institutional/clinical Psoriasis: Department of Skin and STD, Karnataka Institute of Medical Sciences (Hubli, India) and Department of Dermatology, Navodaya Medical College (Raichur, India) Normal: Department of Computer Science, Rani Channamma University (Belagavi, India).	SE: 0.75 SP: 0.67 Precision: 0.60 Acc: 0.70	N/A
Foot ulcer/onychomycosis					
Han et al., 2018	Classification Binary: onychomycosis versus nononychomycosis	Used CNN (ResNet152) to select hand and foot images R-CNN (VGG16) to select nail parts Ensemble method using two CNN (ResNet152 + VGG19) for feature extraction and feedforward neural network for classification	Institutional/dermoscopy Clinical images obtained from four hospitals Trained with Asan dataset Validated with a dataset from Inje University, Hallym University, and Seoul National University	SE/SP/AUC B1 Dataset 96.0/94.7/0.98 B2 Dataset 82.7/96.7/0.95 C Dataset 82.3/79.3/0.93 D Dataset 87.7/69.3/0.82	The clinical photographs used in dermatology are not standardized in terms of image composition. Additional medical photographs may be required for accurate medical diagnoses, and retrieving sufficient numbers of such images may be difficult or even impossible in practical terms
Goyal et al., 2017	Segmentation Automatic segmentation of ulcer and surrounding skin	FCN-AlexNet and FCN-VGG16	Institutional/clinical DFU dataset was collected over a five period at the Lancashire Teaching Hospitals	Dice Coeff: 0.794 for ulcer region, 0.851 for the surrounding skin region, and 0.899 for a combination of both regions	N/A

(*continued*)

Table 1. Continued

Author	Objective	Model Tested	Dataset	Model Performance	Limitations/Comments
Goyal and Hassanpour, 2020[16]	Detection Automatic DFU detection on the DFU challenge dataset	EfficientDet	Public/clinical DFUC2020 provided participants with a comprehensive dataset consisting of 2,000 images for training and 2,000 images for testing	Best test average precision: 53.7 (EfficientDet-D7)	N/A
Cassidy et al., 2021	Classification Automatic DFU detection on the DFU challenge dataset	YOLO, EfficientDet, FRCNN (resnet, inception resnet)	Public/clinical DFUC2020 provided participants with a comprehensive dataset consisting of 2,000 images for training and 2,000 images for testing	Best results Recall: 0.7554 (Inception, version 2 ResNet10 Precision: 0.6919 (EfficientDet) F1-score: 0.6929 (EfficientDet) mAP: 0.6596 (R-FCN)	Authors acknowledge that there is a bias in the dataset, given that the vast majority of subjects are white.
Yap et al., 2021	Classification Automatic DFU detection on the DFU challenge dataset	Faster R−CNN and an ensemble method, YOLOv3, YOLOv5, and EfficientDet	Public/clinical DFUC2020 provided participants with a comprehensive dataset consisting of 2,000 images for training and 2,000 images for testing	mAP: 0.6940 F1-score: 0.7434	N/A
Brüngel and Friedrich, 2021	Classification Automatic DFU detection on the DFU challenge dataset	DETR and YOLOv5	Public/clinical DFUC2020 provided participants with a comprehensive dataset consisting of 2,000 images for training and 2,000 images for testing	DETR F1-score: 0.7355 mAP: 0.7284 YOLOv5 F1-score: 0.7302 mAP: 0.6752 YOLOv5 with TTA F1-score: 0.7351 mAP: 0.7080	N/A
Galdran et al., 2022[17]	Classification Automatic DFU detection on the DFU challenge dataset	Big Image Transfer (BiT), EfficientNet, Vision Transformers, Data-efficient Image Transformers	Public/clinical DFUC2021 dataset	BiT-ResNeXt50 (best result) F1-score: 61.53 AUC: 88.49 Recall: 65.59 Precision: 60.53	N/A
Other					
Maron et al., 2019	Classification Multiclass: five disease classes (AK, intraepithelial carcinoma, benign keratosis, melanocytic nevi, and melanoma)	ResNet50	Public/dermoscopy Trained with ISIC image archive and HAM1000 dataset Tested on biopsy-verified images from HAM1000 dataset	Primary endpoint SE: 74.4% SP: 91.3% Secondary endpoint SE: 56.5% SP: 98.8%	Lack of patient information for AI model algorithms performance would be worse on an entirely external dataset of images.
Zhao et al., 2021	Classification Rosacea, acne, seborrheic dermatitis, and eczema	ResNet50	Institutional/clinical 24,736 photos comprising 18,647 photos of patients with rosacea and 6,089 photos of patients with other skin diseases such as acne, facial seborrheic dermatitis, and eczema	Rosacea Detection Acc: 0.914 Precision: 0.898 AUC: 0.972 Rosacea versus Acne Acc: 0.931 Precision: 0.893 Rosacea versus seborrheic dermatitis and eczema Acc: 0.757 Precision: 0.667	One single dermoscopic image covers only a small proportion of the whole lesion, which hardly represents all the clinical characteristics of the disease comprehensively. Integrating different types of images (clinical, dermoscopic, histopathological) could improve performance.
Aggarwal, 2019	Classification Acne, atopic dermatitis, impetigo, psoriasis, and rosacea	Inception, version 3	Public/clinical Open-source dermatological images captured through DermNet, Dermatology Atlas, Hellenic Dermatological Atlas, and Google images	Average across 5 diseases SE: 0.653 ± 0.045 SP: 0.913 ± 0.027 PPV: 0.660 ± 0.079 NPV: 0.913 ± 0.011 MCC: 0.569 ± 0.074 F1-score: 0.655 ± 0.057	Symptoms such as itching, pain, and other clinical symptoms are absent in the image analysis, which can help the dermatologist in diagnosing the disease
Liu et al., 2020	Classification Multiclass: 26 disease classes (common skin conditions, representing roughly 80% of the volume of skin conditions seen in a primary care setting)	Inception, version 4	Institutional/dermoscopy 16,114 deidentified cases (photographs and clinical data) from a teledermatology practice serving 17 sites	Validation Set A: Top 1: SE: 0.48, PPV: 0.96 Top-3: SE: 0.88, PPV: 0.69 Validation set B: Top 1: SE: 0.57, PPV: 0.95 Top 3: SE: 0.92, PPV: 0.76	Dataset was deidentified, and only structured meta-data was available, which loses information compared with free text clinical notes or an in-person examination

(*continued*)

Table 1. Continued

Author	Objective	Model Tested	Dataset	Model Performance	Limitations/Comments
Haenssle et al., 2020	Classification Multiclass: 10 disease classes (nevus, angioma/ angiokeratoma, SK, dermatofibroma, solar lentigo, AK, Bowen's disease, melanoma, BCC, and SCC)	CNN architecture based on inception, version 4	Public/dermoscopy Tested on MSK-1 dataset (1,100 images) and the ISIC 2018 challenge 16 dataset (1,511 images)	SE: 95.0% SP: 76.7% AUC: 0.918	The test set did not include some other benign (e.g., viral warts), malignant (e.g., Merkel cell carcinoma), or inflammatory (e.g., clear cell acanthoma) skin lesions. Therefore, our results should not be generalized to a large prospective patient population. Dermoscopic images were mostly of patients with a Caucasian genetic background and may not provide comparable results in a population of nonwhite skin types.
Sun et al., 2016	Classification Multiclass: 198 categories	CaffeNet: CNN model pretrained on ImageNet VGGNet	Public/dermoscopy and Clinical SD-198, which contains 198 different diseases from different types of eczema, acne, and various cancerous conditions. There are 6,584 images in total	CaffeNet: 46.69% VGGNet: 50.27%	The dataset shows an imbalance among different categories. Authors tried to collect the same number of samples, whereas some diseases rarely appear in real life
Thomsen et al., 2020	Classification Acne, rosacea, psoriasis, eczema, and cutaneous	VGG 16	Institutional/clinical A total of 19,641 images were provided from the local skin image database of the Department of Dermatology, Aarhus University Hospital (Aarhus, Denmark)	Acne versus Rosacea: SE: 85.42% SP: 89.53% Cutaneous versus Eczema: SE: 74.29% SP: 84.09% Psoriasis versus Eczema: SE: 81.79% SP: 73.57%	One limitation is racial bias, as the data source consisted primarily of Patients with Fitzpatrick skin type II and III. Concerns have been raised about racial bias in CAD in dermatology because databases used for machine learning have historically had an overrepresentation of Caucasian data
Cho et al., 2020	Classification Binary: malignant versus benign lip disorders	Inception-ResNet, version 2	Institutional/dermoscopy Image label split: a total of 1,629 SNUH images (743 malignant and 886 benign) for the training set. The remaining 344 SNUH images (110 malignant and 234 benign) were used as the testing set, along with 281 images (57 malignant and 224 benign) from Seoul National University Bundang Hospital (225 images) and SMG-SNU Boramae Medical Center	344 image set: AUC: 0.827 SE: 0.755 SP: 0.803 281 image set: AUC: 0.774 SE: 0.702 SP: 0.759	Limitations: The algorithm was used to classify binary responses of the diseases and not the likelihood rating from the participants. Most of the images used in this study were of Asian people. The diversity of the diagnoses from the external data of the two affiliated hospitals was lower than that of the training set. The dataset was small compared with those used in previous studies. It is difficult to obtain high-quality, diagnosis-annotated lip images, but these obstacles can be overcome if more appropriate images become available in the future.
Binol et al., 2020	Segmentation/detection Automatically identify rosacea lesions from facial images	Inception-ResNet, version 2, and ResNet-101	Institutional/clinical Images used in this study were captured at the Ohio State University Division of Dermatology	Dice Coeff Inception-ResNet, version 2: 89.8 ± 2.6 % ResNet-101: 87.8 ± 2.4 %	N/A
Kawahara and Hamarneh, 2016	Classification 10 classes: AK, BCC, melanocytic nevus/mole, SCC, SK, IEC, PYO, hemangioma (VSC), DF, and malignant melanoma.	A hybrid of the pretrained AlexNet architecture for early network layers and additional untrained layers for later network layers that learn only from skin images.	Public/dermoscopy Dermofit Image Library: 1,300 skin lesion images from 10 classes	Validation Acc: 0.781 Test Acc: 0.795	N/A

(*continued*)

Table 1. Continued

Author	Objective	Model Tested	Dataset	Model Performance	Limitations/Comments
Wu et al., 2020	Classification Psoriasis, eczema, and atopic dermatitis	EfficientNet-b4	Public/clinical Clinical images from the Department of Dermatology, The Second Xiangya Hospital, Central South University (Xiangya, China)	Overall Acc: 95.80 ± 0.09 % SE: 94.40 ± 0.12 % SP: 97.20 ± 0.06 % Psoriasis Acc: 89.46% SE: 91.4% SP: 95.48% atopic dermatitis and eczema Acc: 92.57% SE: 94.56% SP: 94.41%	N/A

Abbreviations: Acc., accuracy; AI, artificial intelligence; AK, actinic keratosis; AUC, area under the curve; BCC, basal cell carcinoma; CAD, computer aided diagnostic; CNN, convolutional neural network; Coeff, coefficient; DF, dermatofibroma; DFU, diabetic foot ulcer; DI, dice coefficient; dsc, dermatoscopic; IEC, intraepithelial carcinoma; ISBI, international symposium on biomedical imaging; ISIC, International Skin Imaging Collaboration; JA, jaccard index; mAP, mean average precision; MCC, Matthews correlation coefficient; N/A, not applicable; NPV, negative predictive value; PPV, positive predictive value; PYO, pyogenic granuloma; SCC, squamous cell carcinoma; SE, standard error; SK, seborrhoeic keratosis; SMG-SNU, Seoul Metropolitan Goverment-Seoul National University; SNUH, Seoul National University Hospital; SP, specificity; SVM, support vector machine; TTA, test-time augmentation.

[7]Bi L, Kim J, Ahn E, Feng D. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv 2017.

[8]Li KM, Li EC. Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks. arXiv 2018.

[9]Rezvantalab A, Safigholi H, Karimijeshni S. Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. arXiv 2018.

[10]Chang H. Skin cancer reorganization and classification with deep neural network. arXiv 2017.

[11]Mirunalini P, Chandrabose A, Gokul V, Jaisakthi S. Deep learning for skin lesion classification. arXiv 2017.

[12]Murphree DH, Ngufor C. Transfer learning for melanoma detection: participation in ISIC 2017 skin lesion classification challenge. arXiv 2017.

[13]Vasconcelos CN, Vasconcelos BN. Convolutional neural network committees for melanoma classification with classical and expert knowledge based image transforms data augmentation. arXiv 2017.

[14]Sousa RT, de Moraes LV. Araguaia medical vision lab at ISIC 2017 skin lesion classification challenge. arXiv 2017.

[15]Yang X, Zeng Z, Yeo SY, Tan C, Tey HL, Su Y. A novel multi-task deep learning model for skin lesion segmentation and classification. arXiv 2017.

[16]Goyal M, Hassanpour S. A refined deep learning architecture for diabetic foot ulcers detection. arXiv 2020.

[17]Galdran A, Carneiro G, Ballester MAG. Convolutional nets versus vision transformers for diabetic foot ulcer classification. arXiv 2022.

for training CNNs to avoid bias and overestimation of model performance.

Data quality and imbalance. The quality of the images can be a cause for concern, especially with nonpublic institutional datasets, because clinical image quality can vary depending on the device and the operator capturing the images. For dermoscopic images, the images are taken with a designated device, and thus the quality may not vary as much. Quality control of images in large datasets is a challenging task. This issue is compounded by overall lack of large image repositories in dermatology; hence, we note that the largest body of literature is in the melanoma binary classification and diabetic foot ulcer models, given the availability of standardized and publicly available datasets. Increasing the diversity of images in datasets and the development of tools to assess image quality and remove duplicate data are solutions needed to improve model development in the future.

Generalizability of models

Although these studies show a potential use of AI models in dermatology, it should be noted that the majority of papers are largely proof of concept, trained, and tested on retrospective datasets. The limitation in generalizability can be broken down into three categories: lack of datasets in general, lack of diversity in datasets, and lack of patient information. The barriers to generalizability would be the data imbalance across age, sex, ethnicity, skin tone, disease type, and disease prevalence, which if not sufficiently addressed could lead to poor performance of the models when tested outside of their training and test population. For reference, image label splits are noted in Table 2 for standardized, publicly available datasets for comparison with disease prevalence in real-world clinical settings.

One study reported that several ML algorithms may underperform on images from patients with skin of color because the datasets used to train these models such as the ISIC challenge archive have been collected heavily from fair-skinned patients in the United States, Europe, and Australia (Adamson and Smith, 2018). A case study in Uganda showed that only 17% of the images from Fitzpatrick 6 skin (black–dark) type were correctly diagnosed dermatological conditions through First Derm's Skin Image Search algorithm, indicating that the model was mainly trained on Caucasian skin types (Kamulegeya et al., 2019[5]). Similarly, Han et al. (2020a) and Winkler et al. (2019) acknowledged the validation of one race (Asian, Caucasian) in one region (South Korea, Germany). Haenssle et al. (2020) stated that their dataset did not include some other benign, malignant, or inflammatory skin lesions and that the dataset consisted of images from the Caucasian genetic background. Together, these results show that the models will likely not generalize across nonwhite skin types and populations with skin lesion types not included in the dataset used to construct the tested

[5] Kamulegeya LH, Okello M, Bwanika JM, Musinguzi D, Lubega W, Rusoke D, et al. Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. bioRxiv 2019.

Table 2. List of Publicly Available Datasets

Name of Dataset	Dataset Description	Access
ISIC challenge 2016 (melanoma)	Task 1: Lesion Segmentation Training Data: 900 dermoscopic lesion images in JPEG format, with EXIF data stripped Training Ground Truth: 900 binary mask images in PNG format Test Data: 379 images of the same format as the training data Task 2: Detection and Localization of Visual Dermoscopic Features/Patterns Training Data: 807 lesion images in JPEG format and 807 corresponding superpixel masks in PNG format Training Ground Truth: 807 dermoscopic feature files in JSON format Test Data: 335 images of the exact same format as the training data Task 3: Disease Classification Training Data: 900 lesion images in JPEG format Training Ground Truth: 900 entries of gold standard malignant status Test Data: 379 images of the exact same format as the training data Image Label Split: Task 3 (727 benign, 173 malignant)	https://challenge.isic-archive.com/data/
ISIC challenge 2017 (melanoma)	For all three tasks: Training Dataset: 2,000 lesion images in JPEG format and 2,000 corresponding superpixel masks in PNG format, with EXIF data stripped Training ground truth: 2,000 binary mask images in PNG format, 2,000 dermoscopic feature files in JSON format, 2,000 entries of gold standard lesion diagnosis Validation Dataset: 150 images Test dataset: 600 images Image Label Split: Melanoma 374, seborrheic keratosis 254, other (benign): 1,372	https://challenge.isic-archive.com/data/
ISIC Challenge 2018 (melanoma)	For Tasks 1 and 2: Training Dataset: 2,594 images and 12,970 corresponding ground truth response masks (five for each image). Validation Dataset: 100 images Test dataset: 1,000 images For Task 3 (HAM10000 Dataset): Training Dataset: 10,015 images and 1 ground truth response CSV file (containing one header row and 10,015 corresponding response rows). 10,015 entries grouping each lesion by image and diagnosis confirm the type. Training ground truth: 2,000 binary mask images in PNG format, 2,000 dermoscopic feature files in JSON format, and 2,000 entries of gold standard lesion diagnosis Validation Dataset: 193 images Test dataset: 1,512 images Image Label Split: Actinic keratoses and intraepithelial carcinoma/Bowen's disease (327 images), basal cell carcinoma (514 images), benign keratosis-like lesions (1,099 images), dermatofibroma (115 images), melanoma (1,113 images), melanocytic nevi (6,705 images), and vascular lesions (142 images)	ISIC Dataset: https://challenge.isic-archive.com/data/HAM10000 Dataset: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T
ISIC challenge 2019 (melanoma)	Training Set: 25,331 JPEG images of skin lesions and metadata entries of age, sex, and general anatomic site with gold standard lesion diagnosis Test Set: 8,238 JPEG images of skin lesions and metadata entries of age, sex, and general anatomic site. Image Label Split: Actinic keratoses and intraepithelial carcinoma/Bowen's disease (867 images), basal cell carcinoma (3,323 images), benign keratosis–like lesions (2,624 images), dermatofibroma (239 images), melanoma (4,522 images), melanocytic nevi (12,875 images), and vascular lesions (253 images)	https://challenge.isic-archive.com/data/
ISIC challenge 2020 (melanoma)	Training Set: 33,126 DICOM images with embedded metadata and metadata entries of patient ID, sex, age, and general anatomic site with gold standard lesion diagnoses. Test Set: 10,982 DICOM images with embedded metadata and metadata entries of patient ID, sex, age, and general anatomic site. Image Label Split: Benign keratosis-like lesions (37 images), Lentigo (44 images), solar lentigo (7 images), melanoma (584 images), melanocytic nevi (5193 images), seborrheic keratoses (135 images) and other/unknown (benign) (27,124 images)	https://challenge.isic-archive.com/data/
MED—NODE Database	Image Label Split: 100 dermoscopic images: 80 melanomas and 20 nevi 100 nondermoscopic images: 80 melanomas and 20 nevi	https://skinclass.de/mclass/
PH2 Database	Image Label Split: 200 dermoscopic images of melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 melanomas	https://www.fc.up.pt/addi/ph2%20database.html
DermIS Database	Image Label Split: 43 macroscopic photographs with lesions diagnosed as melanoma and 26 diagnosed as nonmelanoma	http://www.dermis.net
DermQuest	Image Label Split: 76 images of melanoma lesions and 61 images of nonmelanoma lesions	http://www.dermquest.com
Interactive Atlas of Dermoscopy	The dataset includes over 2,000 clinical and dermoscopy color images, along with corresponding structured metadata Image Label Split: basal cell carcinoma (42 images), nevi (575 images), dermatofibroma (20 images), lentigo (24 images), melanoma (268 images), miscellaneous (8 images), seborrheic keratoses (45 images), and vascular lesion (29 images) Only 1,011 labels are shown in the dataset	https://derm.cs.sfu.ca/

(*continued*)

Table 2. Continued

Name of Dataset	Dataset Description	Access
DFUC 2020 Dataset	4,000 images, with 2,000 used for the training set and 2,000 used for the testing set. An additional 200 images were used for sanity checking. The training set consists of DFU images only, and the testing set comprised images of DFU and other foot/skin conditions and images of healthy feet. The dataset is heterogeneous, with aspects such as distance, angle, orientation, lighting, focus, and the presence of background objects all varying between photographs. Image Label Split: Not reported unless requested	https://dfu-challenge.github.io/dfuc2020.html
DFUC 2021 Dataset	15,683 DFU patches, with 5,955 training, 5,734 for testing, and 3,994 unlabeled DFU patches Image Label Split: Training set (2,555 infections only, 227 ischemia only, 621 both infection and ischemia, and 2,552 without ischemia and infection)	https://dfu-challenge.github.io/dfuc2021.html

Abbreviations: DFU, diabetic foot ulcer; ISIC, International Skin Imaging Collaboration.

models. To solve these challenges, studies deploying models for prospective validation in real-world settings in which the models will be used are needed. Rigorous validation of models in real-world settings, with training and test data mirroring pretest probability of disease conditions and demographics, will help in generalizability. It should be noted that in current literature, there is a lack of calibration metrics for these AI models. If disease prevalence in the population is known, the model threshold may need to be altered to create a most favorable outcome of data that can meaningfully inform the clinical decision process.

Importance of true labels and ground truth

In larger datasets (MNIST, CIFAR, and ImageNet), deep learning models are able to generalize from training data when true labels far outnumber the incorrect labels (Rolnick et al., 2017[6]). However, in medical datasets, because of the typical smaller sample sizes, it is unclear whether this holds true. For example, neural network training for true melanoma detection from pigmented lesion biopsies by dermatologists is only 9.60 (95% confidence interval = 6.97–13.41) by meta-analysis (Petty et al., 2020). This highlights the importance of using histopathological reports as ground truth for important tasks such as melanoma detection, given that incorrect labels by experts can dilute the dataset, hence creating an inferior model performance. Quantification tasks pose unique challenges such as determining ground truth when there is inter-rater variability from multiple experts, especially in quantification tasks such as inflammatory (urticaria, eczema, etc.) or pigmentation (melasma, post-inflammatory hyperpigmentation) tasks.

Role of clinical information

Most current models are only trained with skin images without consideration of other clinical information related to the patients. Given that physicians usually make clinical decisions with additional information other than imaging, such as with chart reviews, adding this information to the deep learning model could lead to better classification performance. Haenssle et al. (2020) showed that with the addition of the clinical information, there is an increase in the sensitivity of dermatologists' management decisions (89.0–94.1%) and the sensitivity and specificity of diagnostic performance (sensitivity of 83.8–90.6%, specificity of 77.6–82.4%). Thus, it is predicted that deep learning models can benefit from the inclusion of patient metadata.

AI versus human performance

Several studies show comparable diagnostic classification results of AI with those of human experts. However, several algorithms suffer from poor generalizability because of the variable performance of the models when tested outside its experimental conditions (Du-Harpur et al., 2020; Gomolin et al., 2020). This leads to cases of faulty AI, which can have a detrimental impact on the trust and promise that researchers and clinicians have for AI in the realm of dermatology.

Although AI-based classification systems cannot replace human experts, they can cooperate with experts and empower them to make accurate skin diagnoses (Garg et al., 2005, Han et al., 2020a; Hekler et al., 2019). For example, with their CNN model trained on over 200,000 images from four datasets that were further validated on two external datasets, Han et al. (2020b) reported that their model was able to improve the top one accuracy of four dermatologists by 7% in multiclass classification of 134 skin conditions and increase the sensitivity and specificity of malignancy prediction of 47 dermatologists/dermatology residents by 12 and 1%, respectively. Hekler et al. (2019) trained a CNN model using over 11,000 dermoscopic images to perform multiclass classification of five skin conditions and found that the mean combined AI–human accuracy was 83%, which was 1.4% higher than AI alone and 40% higher than experts alone.

AI cannot only assist dermatologists directly but could also be helpful in triage and referral workflow. One study developed a risk-aware neural network model augmented with

Table 3. Types of Datasets

		Types of Images	
		Clinical	Dermoscopy
Data Availability	Open source	[4][20][50–54][57][60][65]	[1,2][5–13][15–20][23–33][35–37][39–41][55][59,60][64]
	Institutional	[3][14][38][42][44][46,47][49][56][61][63]	[3][14][16][19][21,22][34][43][45][48][58][62]

[6] Rolnick D, Veit A, Belongie S, Shavit N. Deep learning is robust to massive label noise. arXiv 2017.

Bayesian deep networks that showed a 90% prediction accuracy in the diagnosis prediction of seven types of skin lesions and made referrals to experts for only 35% of the tested cases (Mobiny et al., 2019). Another study evaluated the impact of AI in assisting primary care providers to diagnose skin conditions (Jain et al., 2021). A total of 40 board-certified clinicians were tasked with diagnosing over 1,000 cases with and without AI assistance, and their results were compared with the reference diagnoses made by dermatologists. The study showed that diagnostic agreement for the primary care physicians increased by 10% and that for nurse practitioners increased by 12% with AI assistance.

Regulatory pathway for approval

The Food and Drug Administration (FDA) is the regulatory entity for approval of any models, typically using the Software as a Medical Device (SaMD) 501K regulatory pathway. FDA has different marketing pathways further explored at https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/how-study-and-market-your-device. Most models are marketed as class II (moderate risk), including those that are CDS tools. Devices that make a definitive diagnosis are often class III (highest risk). The more impact the software has on the healthcare diagnosis/treatment decision, the higher the class attributed to it, a concept that is further explored in the IMDRF software as a medical device risk framework (International medical device regulators, 2014). For approval, the FDA is currently piloting a program for precertification. The proposed concept entails that an FDA review will change on the basis of the risk level of the device; whether it is an initial product review, major change, or minor change; and depending on the organizational experience. All organizations would have to undergo an organizational excellence review to use this program (U.S. Food & Drug Administration, 2021a).

Traditionally, after FDA approval and before marketing, the SaMD must be locked, prohibitive to the adaptive nature of AI/ML software. A discussion paper proposing a novel framework outlining when to submit SaMD modifications can be found (U.S. Food & Drug Administration, 2019). An FDA database (U.S. Food & Drug Administration, 2021b) may be useful to review a current list of approved AI/ML devices.

CONCLUSION AND FUTURE DIRECTIONS

Deep learning has immense potential in dermatology as an assistive diagnostic tool for skin diseases, with promising value in assisting diagnostic and disease quantification tasks. Clinical use spans clinical care, teledermatology, triaging care, and clinical trials among others. The most pressing challenge preventing AI from being more widely used in dermatology is the lack of diversity in datasets and generalizability studies. Working together with physicians and healthcare providers, these AI algorithms can provide more accurate diagnoses and better care, reduce labor costs and workload, and benefit the healthcare industry overall.

ORCIDs

Hyeon K. Jeong: http://orcid.org/0000-0001-6680-2012
Christine Park: http://orcid.org/0000-0002-0066-366X
Ricardo Henao: http://orcid.org/0000-0003-4980-845X
Meenal Kheterpal: https://orcid.org/0000-0002-0460-6400

AUTHOR CONTRIBUTIONS

Conceptualization: HKJ, CP, MK; Funding Acquisition: MK; Project Administration: MK; Supervision: RH, MK; Visualization: HKJ, CP; Writing - Original Draft Preparation: HKJ; Writing - Review and Editing: HKJ, CP, RH, MK

CONFLICT OF INTEREST

The authors state no conflict of interest

REFERENCES

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

Aggarwal SLP. Data augmentation in dermatology image recognition using machine learning. Skin Res Technol 2019;25:815–20.

Araújo RL, Ricardo de Andrade LR, Rodrigues JJ, Silva RR. Automatic segmentation of melanoma skin cancer using deep learning. Paper presented at: 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM). 1–2 March 2021; Shenzhen, China.

Arunkumar TR, Jayanna HS. A novel light weight approach for identification of psoriasis affected skin lesion using deep learning. J Phys Conf Ser 2021;2062:012017.

Ashraf H, Waris A, Ghafoor MF, Gilani SO, Niazi IK. Melanoma segmentation using deep learning with test-time augmentations and conditional random fields. Sci Rep 2022;12:3948.

Binol H, Plotner A, Sopkovich J, Kaffenberger B, Niazi MKK, Gurcan MN. Ros-NET: a deep convolutional neural network for automatic identification of rosacea lesions. Skin Res Technol 2020;26:413–21.

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 2019a;111:148–54.

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.

Brüngel R, Friedrich CM. DETR and YOLOv5: exploring performance and self-training for diabetic foot ulcer detection. A paper presented at: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). 7–9 June 2021; Aveiro, Portugal.

Carin L, Pencina MJ. On deep learning for medical image analysis. JAMA 2018;320:1192–3.

Cassidy B, Kendrick C, Brodzicki A, Jaworek-Korjakowska J, Yap MH. Analysis of the ISIC image datasets: usage, benchmarks and recommendations. Med Image Anal 2022;75:102305.

Cassidy B, Reeves ND, Pappachan JM, Gillespie D, O'Shea C, Rajbhandari S, et al. The DFUC 2020 dataset: analysis towards diabetic foot ulcer detection. touchREV Endocrinol 2021;17:5–11.

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. Br J Dermatol 2020;182:1388–94.

Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). A paper presented at: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). 4–7 April 2018; Washington, DC.

Dash M, Londhe ND, Ghosh S, Semwal A, Sonawane RS. PsLSNet: automated psoriasis skin lesion segmentation using modified U-Net-based fully convolutional network. Biomed Signal Proc 2019;52:226–37.

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. A paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition. 20–25 June 2009; Miami, FL.

Deng L. The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process Mag 2012;29:141–2.

Dietterich TG. Ensemble methods in machine learning. In: Multiple Classifier System MCS 2000. Lecture Notes in Computer Science; 2000.

Du-Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. Br J Dermatol 2020;183:423–30.

Ehteshami Bejnordi BE, Balkenhol M, Litjens G, Holland R, Bult P, Karssemeijer N, et al. Automated detection of DCIS in whole-slide H&E

stained breast histopathology images. IEEE Trans Med Imaging 2016;35:2141–50.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Corrigendum: dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;546:686.

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.

Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293:1223–38.

Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021;12:711.

Gomolin A, Netchiporouk E, Gniadecki R, Litvinov IV. Artificial intelligence applications in dermatology: where do we stand? Front Med (Lausanne) 2020;7:100.

Gown AM, Goldstein LC, Barry TS, Kussick SJ, Kandalaft PL, Kim PM, et al. High concordance between immunohistochemistry and fluorescence in situ hybridization testing for HER2 status in breast cancer requires a normalized IHC scoring system. Mod Pathol 2008;21:1271–7.

Goyal M, Oakley A, Bansal P, Dancey D, Yap MH. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. IEEE Access 2019;8:4171–81.

Goyal M, Yap MH, Reeves ND, Rajbhandari S, Spragg J. Fully convolutional networks for diabetic foot ulcer segmentation. A paper presented at: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 5–8 October 2017; Banff, Alabama, Canada.

Guergueb T, Akhloufi MA. Melanoma skin cancer detection using recent deep learning models. Annu Int Conf IEEE Eng Med Biol Soc 2021;2021:3074–7.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.

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, 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. JAMA Dermatol 2020a;156:29–37.

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 2018;13:e0191493.

Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. J Invest Dermatol 2020b;140:1753–61.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Computer Vision Foundation, Las Vegas Valley, NV, 2016:770–8.

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 2019;120:114–21.

International medical device regulators. Software as a medical device. possible framework for risk categorization and corresponding considerations. https://www.imdrf.org/documents/software-medical-device-possible-framework-risk-categorization-and-corresponding-considerations; 2014. (accessed February 10, 2022).

Jafari MH, Nasr-Esfahani E, Karimi N, Soroushmehr SMR, Samavi S, Najarian K. Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma. Int J Comput Assist Radiol Surg 2017;12:1021–30.

Jain A, Way D, Gupta V, Gao Y, de Oliveira Marinho G, Hartford J, et al. Development and assessment of an artificial intelligence–based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. JAMA Netw Open 2021;4:e217249.

Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2:230–43.

Jojoa Acosta MF, Caballero Tovar LY, Garcia-Zapirain MB, Percybrooks WS. Melanoma diagnosis using deep learning techniques on dermatoscopic images. BMC Med Imaging 2021;21:6.

Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020;145:463–9.

Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health Inform 2018;23:538–46.

Kawahara J, Hamarneh G. Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers. In: Wan L, Adeli E, Wang Q, Shi Y, Suk HI, editors. Machine learning in medical imaging. MLMI 2016. Lecture Notes in Computer Science. Heidelberg, Germany: Springer; 2016. p. 164–71.

Kaymak S, Esmaili P, Serener A. Deep learning for two-step classification of malignant pigmented skin lesions. A paper presented at: 2018 14th Symposium on Neural Networks and Applications (NEUREL). 20–21 November 2018; Belgrade, Serbia.

Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172:1122–31.e9.

Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018;24:1716–20.

Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284:574–82.

Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 2020;11:3852.

Li YX, Shen LL. Skin lesion analysis towards melanoma detection using deep learning network. Sensors (Basel) 2018;18:556.

Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer vision – ECCV 2014. ECCV 2014. Lecture notes in computer science. Heidelberg, Germany: Springer; 2014. p. 740–55.

Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26:900–8.

Lopez AR. Giro-i-Nieto X, Burdick J, Marques O. Skin lesion classification from dermoscopic images using deep learning techniques. A paper presented at: 2017 13th IASTED International Conference on Biomedical Engineering (BioMed). 20–21 February 2017; Innsbruck, Austria.

Mahbod A, Schaefer G, Wang CL, Ecker R, Ellinger I. Skin lesion classification using hybrid deep neural networks. A paper presented at: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 12–17 May 2019; Brighton, United Kingdom.

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.

Meienberger N, Anzengruber F, Amruthalingam L, Christen R, Koller T, Maul JT, et al. Observer-independent assessment of psoriasis-affected area using machine learning. J Eur Acad Dermatol Venereol 2020;34:1362–8.

Menegola A, Fornaciali M, Pires R, Bittencourt FV, Avila S, Valle E. Knowledge transfer for melanoma screening with deep learning. A paper presented at: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). 18–21 April 2017; Melbourne, Australia.

Mishra R, Daescu O. Deep learning for skin lesion segmentation. A paper presented at: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 13–16 November 2017; Kansas City, MO.

Mobiny A, Singh A, Van Nguyen H. Risk-aware machine learning classifier for skin lesion diagnosis. J Clin Med 2019;8:1241.

Papadakis M, Paschos A, Manios A, Lehmann P, Manios G, Zirngibl H. Computer-aided clinical image analysis for non-invasive assessment of tumor thickness in cutaneous melanoma. BMC Res Notes 2021;14:232.

Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 2016;35:1240–51.

Petty AJ, Ackerson B, Garza R, Peterson M, Liu B, Green C, et al. Meta-analysis of number needed to treat for diagnosis of melanoma by clinical setting. J Am Acad Dermatol 2020;82:1158–65.

Pomponiu V, Nejati H, Cheung NM. Deepmole: deep neural networks for skin mole lesion classification. A paper presented at: 2016 IEEE International Conference on Image Processing (ICIP). 25–28 September 2016; Phoenix, AZ.

Pour MP, Seker H, Shao L. Automated lesion segmentation and dermoscopic feature segmentation for skin cancer analysis. Annu Int Conf IEEE Eng Med Biol Soc 2017:640–3.

Qaiser T, Mukherjee A, Reddy Pb C, Munugoti SD, Tallam V, Pitkäaho T, et al. HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology 2018;72:227–38.

Raj R, Londhe ND, Sonawane RRS. Automated psoriasis lesion segmentation from unconstrained environment using residual U-Net with transfer learning. Comput Methods Programs Biomed 2021;206:106123.

Sagi O, Rokach L. Ensemble learning: a survey. WIREs Data Mining Knowl Discov 2018;8:e1249.

Salamaa WM, Aly MH. Deep learning design for benign and malignant classification of skin lesions: a new approach. Multimed Tools Appl 2021;80:26795–811.

Shahin AH, Kamal A, Elattar MA. Deep ensemble learning for skin lesion classification from dermoscopic images. A paper presented at: 2018 9th Cairo International Biomedical Engineering Conference (CIBEC). 20–22 December 2018; Cairo, Egypt.

Shorfuzzaman M. An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. Multimedia Syst 2022;28:1309–23.

Sun XX, Yang JF, Sun M, Wang K. A benchmark for automatic visual classification of clinical skin disease images. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer vision - ECCV 2016. ECCV 2016. Lecture Notes in Computer Science. Heidelberg, Germany: Springer; 2016. p. 206–22.

Thomsen K, Christensen AL, Iversen L, Lomholt HB, Winther O. Deep learning for diagnostic binary classification of multiple-lesion skin diseases. Front Med (Lausanne) 2020;7:574329.

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 2019;155:58–65.

Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5:180161.

U.S. Food & Drug Administration. Digital health software precertification (pre-cert) program. https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program; 2021. (accessed February 10, 2022)

U.S. Food & Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf; 2019. (accessed February 10, 2022).

U.S. Food & Drug Administration. Artificial intelligence and machine learning (AI/ML) software as a medical device action plan. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device; 2021. (accessed February 10, 2022))

Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: a brief review. Comput Intell Neurosci 2018;2018:7068349.

Webster DE, Suver C, Doerr M, Mounts E, Domenico L, Petrie T, et al. The Mole Mapper Study, mobile phone skin imaging and melanoma risk data collected using ResearchKit. Sci Data 2017;4:170005.

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.

Wu HJ, Yin H, Chen HP, Sun MY, Liu XQ, Yu YZ, et al. A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. Ann Transl Med 2020;8:581.

Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. Sun database: large-scale scene recognition from abbey to zoo. A paper presented at: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 13–18 June 2010; San Francisco, CA.

Yan K, Wang X, Lu L, Summers RM. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J Med Imaging (Bellingham) 2018;5:036501.

Yang Y, Wang J, Xie F, Liu J, Shu C, Wang Y, et al. A convolutional neural network trained with dermoscopic images of psoriasis performed on par with 230 dermatologists. Comput Biol Med 2021;139:104924.

Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. Exp Dermatol 2018;27:1261–7.

Yap MH, Cassidy B, Pappachan JM, O'Shea C, Gillespie D, Reeves ND. Analysis towards classification of infection and ischaemia of diabetic foot ulcers. A paper presented at: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). 27–30 July 2021; Athens, Greece.

Yu C, Yang S, Kim W, Jung J, Chung KY, Lee SW, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images [published correction appears in PLoS One 2018;13:e0196621] PLoS One 2018;13:e0193321.

Yu LQ, Chen H, Dou Q, Qin J, Heng PA. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging 2017;36:994–1004.

Zhao S, Xie B, Li Y, Zhao X, Kuang Y, Su J, et al. Smart identification of psoriasis by images using convolutional neural networks: a case study in China. J Eur Acad Dermatol Venereol 2020;34:518–24.

Zhao Z, Wu CM, Zhang S, He F, Liu F, Wang B, et al. A novel convolutional neural network for the diagnosis and classification of rosacea: usability study. JMIR Med Inform 2021;9:e23415.