The Possibility of Deep Learning-Based, Computer-Aided Skin Tumor Classifiers

Yasuhiro Fujisawa*, Sae Inoue and Yoshiyuki Nakamura

Department of Dermatology, University of Tsukuba, Tsukuba, Japan

The incidence of skin tumors has steadily increased. Although most are benign and do not affect survival, some of the more malignant skin tumors present a lethal threat if a delay in diagnosis permits them to become advanced. Ideally, an inspection by an expert dermatologist would accurately detect malignant skin tumors in the early stage; however, it is not practical for every single patient to receive intensive screening by dermatologists. To overcome this issue, many studies are ongoing to develop dermatologist-level, computer-aided diagnostics. Whereas, many systems that can classify dermoscopic images at this dermatologist-equivalent level have been reported, a much fewer number of systems that can classify conventional clinical images have been reported thus far. Recently, the introduction of deep-learning technology, a method that automatically extracts a set of representative features for further classification has dramatically improved classification efficacy. This new technology has the potential to improve the computer classification accuracy of conventional clinical images to the level of skilled dermatologists. In this review, this new technology and present development of computer-aided skin tumor classifiers will be summarized.

Keywords: artificial intelligence, deep learning, convolutional neural network, clinical image, dermoscopy, skin tumor classifier
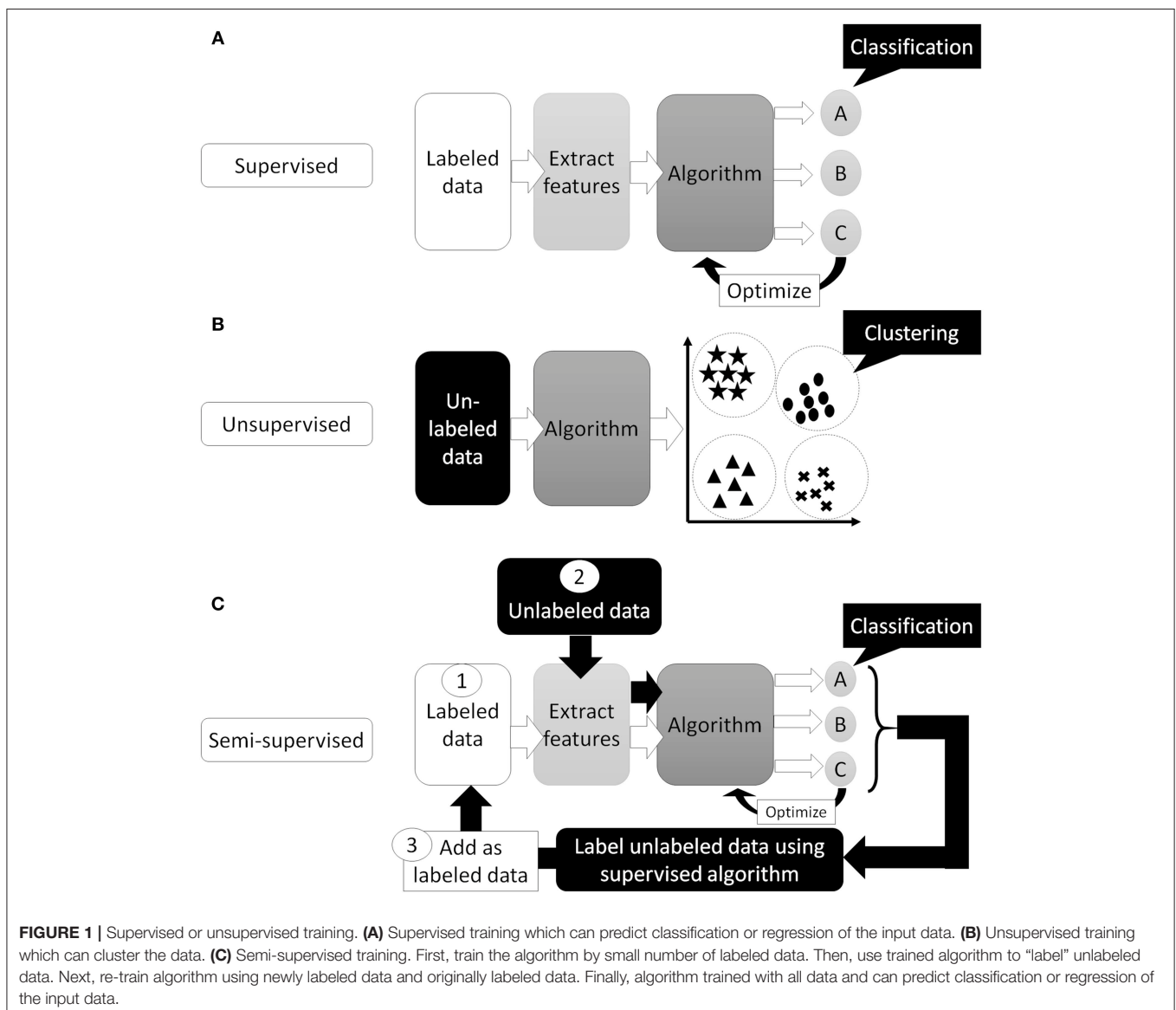
INTRODUCTION

The incidence of skin cancers, including melanoma and non-melanoma skin cancers (NMSC), is globally increasing. In the United States, the incidence of melanoma is reported to be 22.1 per 100,000 people, the number of new yearly melanoma patients is estimated to be more than 63,000, and melanoma is now rated as the 6th most common of all cancers (1). In spite of new therapeutic agents, such as checkpoint and BRAF inhibitors which improve survival of advanced cases, melanomas are still lethal (2–4). On the other hand, most NMSCs, which are responsible for 4.3–5.4 million new cases each year in the United States (5, 6), can be treated simply by surgical removal. Most of these (>90%) are comprised of basal cell carcinomas (BCC) and squamous cell carcinomas (SCC) (7) and skilled dermatologists can detect these by clinical appearance and a tumor magnifying dermatoscope (8). Consequently, most SCCs and BCCs are detected at an early stage and resolved by surgery alone. However, SCC can become lethal when it metastasizes, since few standardized and effective therapies for advanced SCC have been established. Although metastatic BCC is very rare, any delay in diagnosis may allow tumors to become unresectable. Therefore, early detection of all skin cancers, not limited to melanoma, is required to prevent progression of these cancers to advanced stages and reduce skin cancer-related deaths.

Skin tumor screening is one solution for the early detection of skin cancer, but it is not practical for dermatologists to check all patients for skin tumors. In most countries, primary vigilance is maintained through primary care clinics before being referred to dermatologists and, consequently, up to 20% of patients consult at primary care clinics complaining about skin-related symptoms (9–11). A study by Julian et al. reported that 16% of patients with dermatologically related diseases had benign skin tumors, 3.3% had actinic keratosis, and 3% had malignant skin tumors (10), meaning ∼20% of patients who consulted a primary care doctor with skin-related complaints received a tumor diagnosis. Similar results were reported by Kerr et al. (11), showing that 11.4% of studied patients had benign skin tumors with <5% having malignant skin tumors. Although the percentage of patients with malignant skin tumors among patients who consult at primary care clinics is not so high, primary care doctors are under a heavy burden to correctly screen patients who present with skin-related symptoms and determine which patients are to be transferred to the dermatologists. Thus, any device or service that can accurately give the probability of malignancy by analyzing a simple photograph of the tumor would be very helpful for both primary care doctors and their patients. In this context, the development of artificial intelligence (AI) that can classify skin tumor images within seconds, at a skill level similar to trained dermatologists, is an ideal solution for this problem.

MACHINE LEARNING: NECESSITY OF LABELED DATA

Artificial Intelligence (AI) is a term used to describe machine software that can mimic human cognitive functions, such as learning and problem solving (12). Machine learning achieves this via changes in the program algorithm that allow it to complete tasks more efficiently. These changes

FIGURE 1 | Supervised or unsupervised training. **(A)** Supervised training which can predict classification or regression of the input data. **(B)** Unsupervised training which can cluster the data. **(C)** Semi-supervised training. First, train the algorithm by small number of labeled data. Then, use trained algorithm to "label" unlabeled data. Next, re-train algorithm using newly labeled data and originally labeled data. Finally, algorithm trained with all data and can predict classification or regression of the input data.
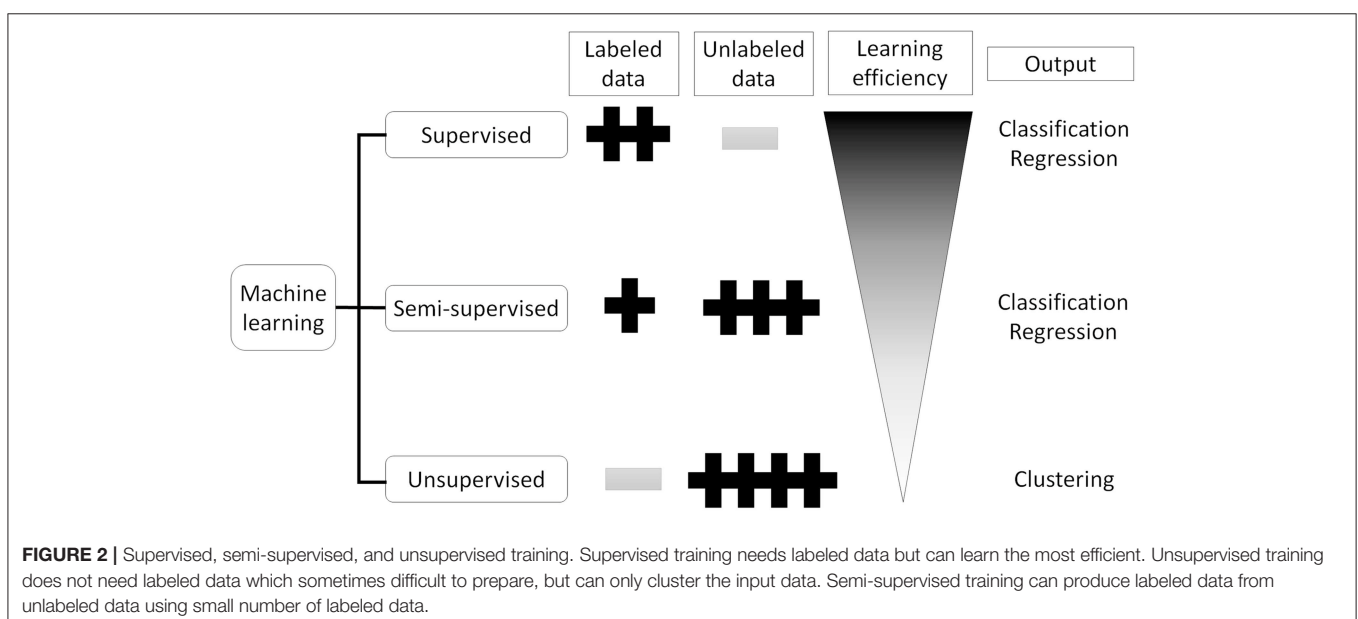
come from training using labeled data (supervised learning method, **Figure 1A**), data without labels (unsupervised learning method, **Figure 1B**), or both (semi-supervised learning method, **Figure 1C**) (13). For the supervised learning method, the program processes data and compares its output with the correct answer (label), adjusting its own parameters so that it can reach the correct result. This process should be repeated for as many training datasets as are available but, to achieve satisfactory efficiency, it requires a certain amount of labeled data to adjust the parameters. Thus, preparing a high enough number of datasets means that this supervised learning method could achieve a high classification efficacy (**Figure 2**). However, preparation of labeled data is often difficult, especially in the medical field. On the other hand, unsupervised learning does not require labeled data but instead uses a large amount of unlabeled data for learning. Where it differs from the supervised method is in the output of this algorithm, which clusters data instead of classifying it (**Figure 1B**). This method is useful for huge amounts of data without labels but, since this algorithm does not know the "answer," the meaning of each cluster in the output needs to be determined. The third method, semi-supervised learning, requires a small amount of labeled data with a large amount of unlabeled data. It functions on the principle that unlabeled data is classified using an algorithm trained with labeled data (**Figure 1C**). These unlabeled data are labeled with categories that were found to have high hit probability as calculated by the algorithm which was trained (supervised learning method) on the small amount of labeled data. These newly labeled data are added to the originally labeled data and supervised learning is then conducted again to re-train the algorithm. This method is useful for huge amounts of unlabeled data that would otherwise require a high cost to label. However, if a small error exists in the initial supervised learning algorithm, that error will be amplified at the end of the procedure. Collectively, the best way to train

AI algorithms is to prepare large amounts of correctly labeled data for supervised learning but this is often difficult for the medical field, since the "gold standard" pathological confirmation of lesion labels requires the excision of "all" lesions, which is not ethical for confirmed, benign lesions.

METHOD OF MACHINE LEARNING FOR IMAGE CLASSIFICATION: BEFORE THE DEEP LEARNING ERA

Developing a computer-aided diagnostic support system for skin cancer diagnosis requires many steps, as reviewed by Masood and Al-Jumaily (14) (**Figure 3**). The first step of the classification process starts by removing irrelevant structures and artifacts in the image (14, 15), such as hair, air bubbles/gel (if dermoscopy images are used), ink markings or reflectance, by using general image filters (16). These irrelevant objects or artifacts may affect the efficacy of border detection and skew the final output. Next, to analyze the internal properties for further analysis, the lesion within the objective tumor area should be separated from the surrounding skin in a procedure called segmentation. As it is not practical to manually define areas for all images, many automatic lesion segmentation systems have been reported (17–19), but this step is still a challenging task for engineers (16).

Next, the important features need to be extracted from the segmented image. Border shapes [asymmetry indices, symmetry axes, or aspect ratios (20)] or color features [average values and standard deviation of the RGB or HSV color channels (21)] are calculated and these values are used for further classification. However, there is a cost for attempting to extract more features, namely more training time, more complex algorithms, less generalization behavior, and less prediction accuracy. Thus, it is important to select only the useful feature values for classification

FIGURE 2 | Supervised, semi-supervised, and unsupervised training. Supervised training needs labeled data but can learn the most efficient. Unsupervised training does not need labeled data which sometimes difficult to prepare, but can only cluster the input data. Semi-supervised training can produce labeled data from unlabeled data using small number of labeled data.
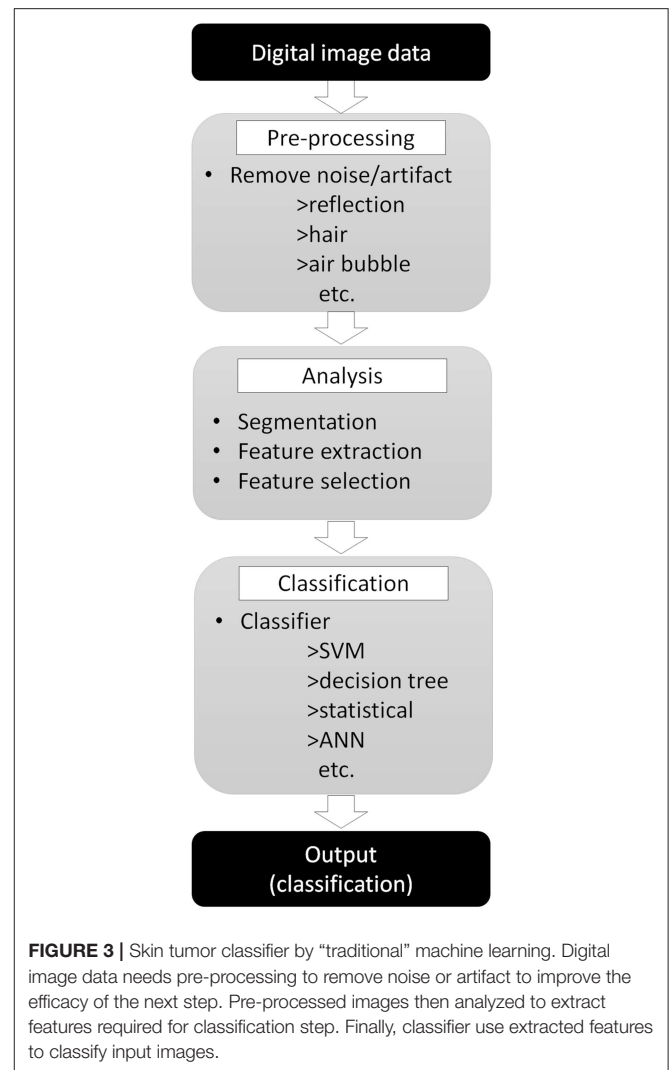
while eliminating less useful ones (feature selection). There are diverse and numerous methods that have been proposed for this feature selection process (22–24).

Finally, the classification algorithm outputs the result using the selected feature values calculated in the previous phase. There are many different algorithms available for this classification task: support vector machine (25), decision tree (26), statistical [logistic regression (27)], or artificial neural network (ANN) (28). Of those, the performance of the support vector machine classifier is reported to be similar or better than other algorithms but, as it can only provide a dichotomous distinction between two classes (e.g., benign or malignant), this algorithm will not work for multi-class sorting with probabilities for each class. ANN, on the other hand, mimics the structure of biological neural networks in the human brain (**Figure 4A**) and can change connectivity between decision nodes (back propagation, **Figure 4B**) so that the network can achieve satisfactory results (28). Many ANN studies have reported on dermoscopic image analysis as (28) it has the ability to derive meaning from data which is too complex for humans to understand. The downside, however, is that ANN requires multiple repetitions of the training data to adjust network connections.

## INTRODUCTION OF DEEP LEARNING TECHNOLOGY WITH CONVOLUTIONAL NEURAL NETWORKS

The image classification machine learning algorithms described above are very complex; they are based on hand-engineered features and are highly dependent on prior knowledge. For example, in some reports, more than 50 different feature values thought to be useful in the classification process, such as color, shape, or border information, were extracted from a single image for the training of the system (29, 30). In the annual ImageNet Large Scale Visual Recognition Challenge (ILSVR) computer vision competition, where 1 to 2 million images of objects are classified into 1,000 categories, a classifier using traditional machine learning had an error rate of 30% (31) compared to humans who logged an error rate of 5.1% (32). This striking gap in accuracy was dramatically reduced in the 2012 ILSVR competition; deep-learning technology using a convolutional neural network (CNN) achieved an error rate of 16.4% while other classifiers using traditional machine learning had an error rate of 26–30% (32). After the introduction of CNN, the error rates in the ILSVR competition dropped rapidly and the error rate in the 2017 competition was below 5%, indicating that CNN classified images more precisely than humans (http://image-net.org/challenges/LSVRC/2017/results).

This new CNN technology can learn and automatically determine what features are important for classification from the training image set. The extraction and selection of the features for classification was a key component of the traditional methods (33), and also the most difficult part. Thus, by using CNN, complicated image pre-processing is no longer necessary to obtain optimal feature values for the image classification. A schematic structure of CNN is shown in **Figure 5A**. In ANN,

FIGURE 3 | Skin tumor classifier by "traditional" machine learning. Digital image data needs pre-processing to remove noise or artifact to improve the efficacy of the next step. Pre-processed images then analyzed to extract features required for classification step. Finally, classifier use extracted features to classify input images.

every node fully connects to the next layer (**Figure 4B**) but, in CNN, each node connects only to some nodes in the next layer (**Figure 5A**). This key feature of CNN can successfully capture the spatial and temporal dependencies in an image through the application of relevant filters (34). In this type of classifier, output values of the feature extractor usually input to a fully connected network and the softmax function finally converts input vectors to real numbers for normalization into a probability distribution (35). As an example, if the input images had 4 different classes, the final CNN layer would have 4 nodes as in **Figure 5A**. But, as the sum of the output of all 4 nodes would not be 1, it would be difficult to interpret the output. However, a softmax function that converts each node's output from 0 to 1 would allow for the components to add up to 1 and result in a final output that can be interpreted as a probability (0–100% probability).

There are many available CNN architectures used in the medical field such as LeNet (36), AlexNet (36), ZFNet (37), VGGNet (38), GoogLeNet (39, 40), ResNet (41), or SENet (42) (**Table 1**). Not only are these architectures free to use, pre-trained models are also available that are commonly trained by
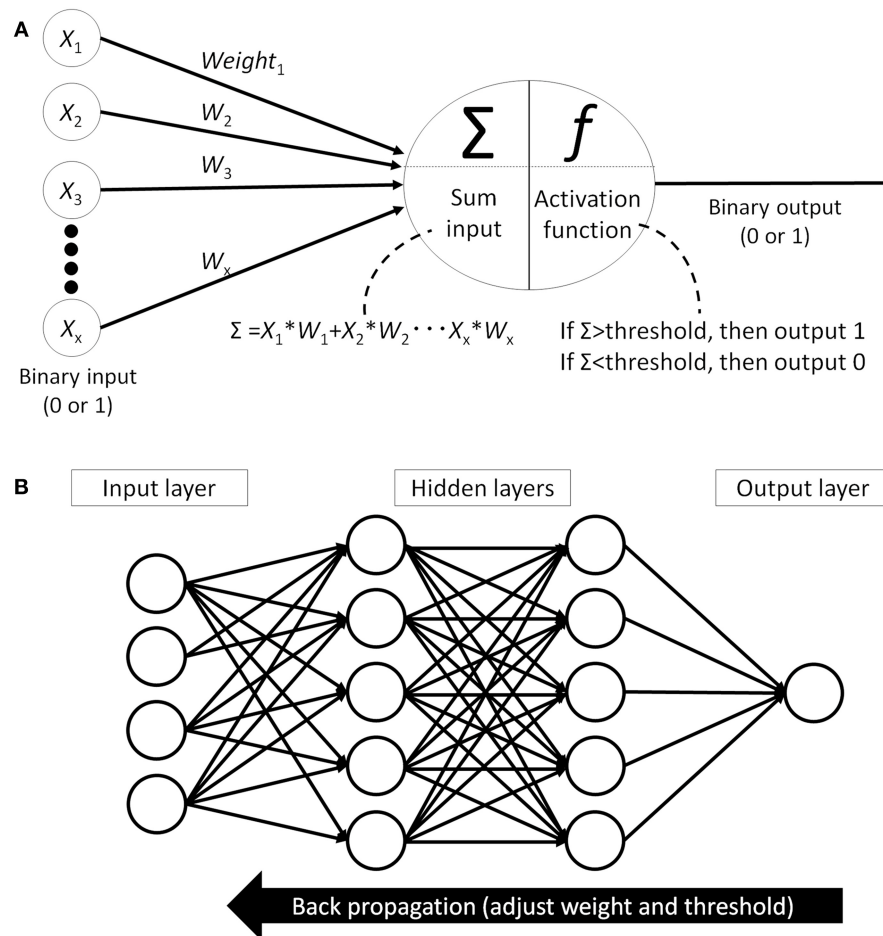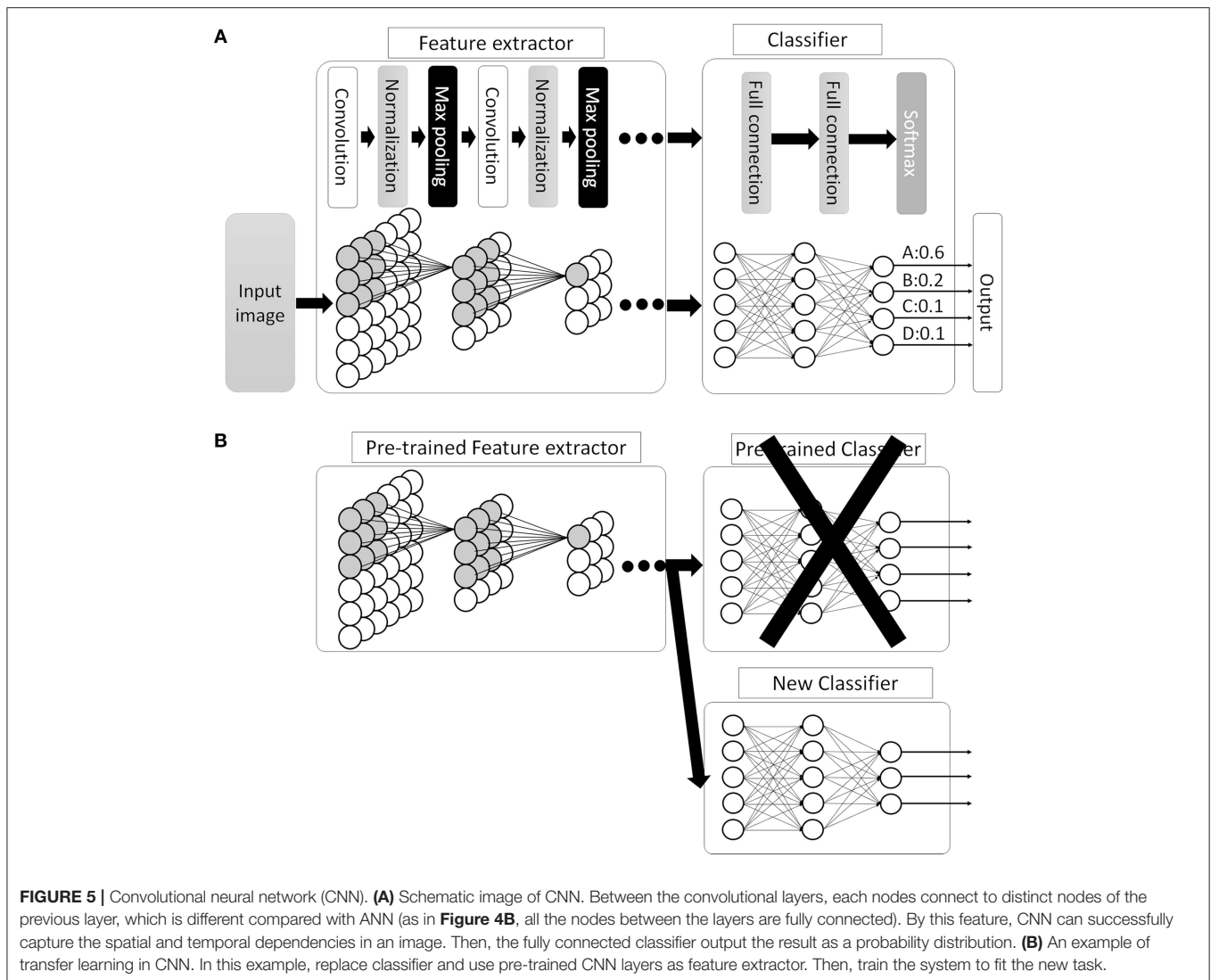
FIGURE 4 | Artificial neural network (ANN). **(A)** Single perceptron model which mimics the structure of biological neural networks in the human brain. Each node receives signal from other nodes (X1, X2… Xx). Add the multiplied values of input and weight (W) and when this sum(Σ) cross the threshold, then this node outputs signal. **(B)** An example of artificial neural network model which has hidden layer between input layer and output layer. All the nodes between the layers are fully connected and each connection has weight. Machine learning is adjusting each weights and thresholds in the network to reach the correct output (back propagation).

the previously mentioned ILSVR2012 dataset, which contains 1.2 million images within 1,000 classes (available at http://image-net.org/download-imageurls). Since the ability to extract image features by pre-trained models is very high, we can use these pre-trained models as a "feature extractor" in a technique called transfer learning (43). One example of transfer learning is shown in **Figure 5B**. Basically, the classifier part is replaced with an untrained classifier appropriate to the new task and the system is trained using a new training image dataset. This method is useful when large numbers of datasets cannot be prepared due to rarity, expense in collection/labeling, or inaccessibility (43). Therefore, transfer learning would be useful for the medical field since it is often difficult to collect images of rare diseases.

Collectively, the introduction of CNN has not only dramatically improved image classification efficacy, it has also made adoption of machine learning and image classification easier and cheaper since most of the initially needed resources are easily accessible.

SKIN TUMOR CLASSIFICATION BY USING DERMOSCOPIC IMAGES

The clinical diagnosis of melanoma is difficult since the morphological characteristics of other pigmented skin lesions may sometimes mimic it. Dermoscopy can magnify the skin and enables clinicians to better evaluate morphological features which are difficult to see with the naked eye. The introduction of dermoscopy has been reported to improve diagnostic sensitivity by 10–30% (44–46). Physicians are usually taught the ABCD-rule (47), Menzie's method (48), 7-point checklist (48), or some other pattern classification methods (49) to distinguish between melanoma and non-melanoma pigmented skin lesions. However, becoming an experienced dermoscopic reader (50, 51) who can score 90% diagnostic sensitivity (proportion of images correctly detected as malignant within all malignant images) and 59% diagnostic specificity (proportion of images correctly detected as benign within all benign images) requires a significant time and training investment

FIGURE 5 | Convolutional neural network (CNN). **(A)** Schematic image of CNN. Between the convolutional layers, each nodes connect to distinct nodes of the previous layer, which is different compared with ANN (as in **Figure 4B**, all the nodes between the layers are fully connected). By this feature, CNN can successfully capture the spatial and temporal dependencies in an image. Then, the fully connected classifier output the result as a probability distribution. **(B)** An example of transfer learning in CNN. In this example, replace classifier and use pre-trained CNN layers as feature extractor. Then, train the system to fit the new task.

(52). Moreover, even after such training, the readings are often complex and subjective.

To make readings more objective and qualitative, as well as support physicians using dermoscopy, many computer-based analyses of dermoscopic images to classify melanomas have been conducted. In a 2009 review by Rajpara et al. (53) that reviewed studies of AI classifiers (12 studies) and dermoscopy (23 studies) published between 1991 and 2002, the melanoma detection sensitivity and specificity of AI classifiers were already similar to that of physicians using dermoscopy; pooled sensitivity, and specificity for AI classifiers and physicians were 91 vs. 88% and 86 vs. 79%, respectively. In a 2013 review (14), 15 new studies on AI classifiers were included and showed sensitivities ranging from 60.7 to 98% and specificities ranging from 72 to 100%, which were similar to the previous 2009 report. Haenssle et al. reported on an AI classifier using the CNN deep learning-based algorithm and compared the classification efficiency against 58 dermatologists (54). Their CNN showed higher accuracy than most dermatologists (sensitivity 88.9 vs.

86.6% and specificity 82.5 vs. 71.3%, respectively) (54). Similar results were recently reported by Tschandl et al. (55) which compared 511 human readers, including 283 board-certified dermatologists, and 139 machine learning algorithms on a classification task consisting of 30 image batches from the test image set. When comparing 37 human experts (>10 years of experience) with the top three machine learning algorithms, the mean number of correctly classified images by humans was 18.78 images per 30 test images, whereas machine learning algorithms scored 25.43, which was statistically higher than human expert readers.

Although AI classifiers seemed to score similar marks as physicians, it is difficult to judge whether AI classifiers have already surpassed physicians or not, since most of these reports were unable to verify results by outside data and biases, such as selection bias of study training/testing data or publication bias (53). However, in spite of such possible biases, development of AI classifiers using dermoscopic images is still an attractive research area.

TABLE 1 | List of CNN architectures.

Architecture	Year	Top-5 error rate at ILSVRC*	References
LeNet	1998	NA	(36)
AlexNet	2012	15.3%	(36)
ZFNet	2013	14.8%	(37)
GoogLeNet	2014	6.67%	(39, 40)
VGG Net	2014	7.3%	(38)
ResNet-50	2015	3.6%	(41)
SENet	2017	2.3%	(42)

ILSVRC: ImageNet Large Scale Visual Recognition Challenge.

CLASSIFICATION OF SKIN TUMOR USING CLINICAL IMAGES

As described above, many highly accurate AI classifiers focusing on melanoma detection using dermoscopic images have been developed. Although conventional machine learning algorithms that require human intervention to extract and select features have the capacity to output a binary result (benign or malignant), accurate diagnoses of multi-class skin diseases are difficult (56). Moreover, even if clinical images are cheap and easy to collect, these clinical photographs are believed to have limited morphologic information that is useful for classification (14). Collectively, outside of a binary "benign or malignant" output, AI classifiers using conventional machine learning algorithms are considered to be inferior at handling clinical images for sorting into multiple classes.
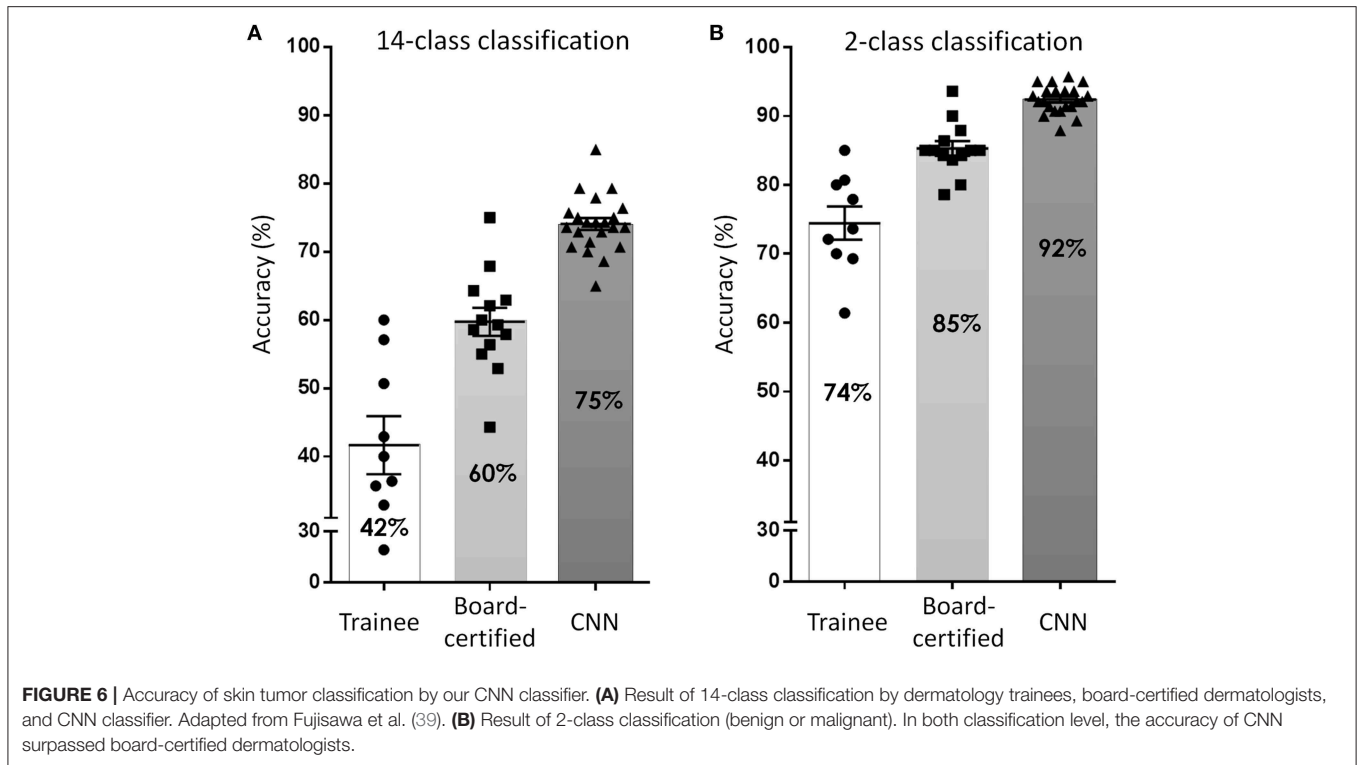
To overcome these issues, deep learning-based CNN classifiers, which surpass the general object classification capability of humans (32), became popular for use in these tasks. Several studies have been published, including from our group (39, 57, 58), since Esteva et al. (40) first reported a dermatologist-equivalent classifier of skin cancers using CNN in 2017. They used 129,450 skin lesion images for the training of a Google Inception v3 CNN architecture, which was pre-trained on the Image Net dataset consisting of 1.28 million images over 1,000 generic object classes. The CNN was fine-tuned to classify skin lesion images by the transfer learning method and was validated on its efficiency of binary classification (benign or malignant). Although the study did not reveal its overall accuracy at skin tumor classification, the CNN surpassed average-level dermatologists in the sensitivity/specificity of classifying epidermal tumors (epidermal cancers and seborrheic keratosis) and melanocytic tumors (melanoma and benign nevi). Han et al. (57) used 15,408 skin lesion images from 12 benign and malignant skin tumors to train a Microsoft ResNet-152 CNN architecture (pre-trained on the same Image Net dataset as above). Similarly, to the Esteva et al. report, they used skin tumor images to fine-tune their CNN which subsequently outperformed 16 dermatologists. They also opened their CNN to the public and it could be externally validated, which was noteworthy. In our study (39), we used only 4,800 skin lesion images from 14 benign and malignant skin tumors to train a GoogLeNet CNN architecture (pre-trained on the same Image Net dataset as above). However, even with less images in the training set, our CNN was more accurate at image classification than 13 board-certified dermatologists and 8 dermatologist trainees (**Figure 6**), reaching a 96.3% sensitivity, and 89.5% specificity in the detection of skin cancer.

Brinker et al. (58) reported an interesting result showing that a CNN trained only with dermoscopic images could classify clinical melanoma images at a similar level to 145 dermatologists. They trained a Microsoft ResNet-50 CNN architecture (pre-trained on the same Image Net dataset as above) using 2,196 melanomas and 18,566 atypical nevi. This study is particularly interesting because this is the first report to show that dermatologist-equivalent tumor image classification was achieved by a CNN that was not trained by clinical images. This study indicates that CNN may benefit from training with dermoscopic images (that have a higher resolution than clinical images) even for low-resolution classification tasks. Another approach is to combine available data for classification as Yap et al. (59). They used 2,917 cases containing both clinical and dermoscopic images and trained a Microsoft ResNet-50 CNN architecture. They showed that a CNN trained with dermoscopic images had higher accuracy than a CNN trained with clinical images. However, when they trained their CNN on combined feature information from dermoscopic and clinical images, the accuracy outperformed single modal CNN, indicating that both clinical and dermoscopic images have distinct classification information. Collectively, the new machine learning algorithm CNN could be a "breakthrough" for developing a multi-class skin tumor classifier, which can accept clinical tumor images.

LIMITATIONS

Several issues remain for the CNN skin tumor classifier to overcome. First, there are no standardized evaluation test datasets to measure the efficacy of CNN classifiers. However, if the test dataset is known in advance, there is a risk of adapting the CNN classifier to the test dataset. Therefore, it might be better to conduct tests by a third-party organization to measure classification efficiency using closed datasets. Second, datasets used to train the CNN are comprised of regionally homogenous images, e.g., our dataset was composed of nearly 100% Asians. In a study by Han et al. (57), they tested using external tumor images (Edinburgh dataset; available from the Edinburgh Dermofit Image Library) to see if their CNN that was trained on Asian tumor images could also classify tumor images from Caucasian patients. As anticipated, both sensitivity and specificity dropped in this case. Third, although CNN requires an increased number of training datasets to improve classification efficiency, rare tumors and subtypes (such as amelanotic melanoma or pigmented basal cell carcinoma) will always mean a scarcity of available images for these diseases. Fourth, the clinical images were less standardized, with varying camera angles, orientations, multiple skin backgrounds, lighting, and even pen markings or rulers included in the photos (60). According to a study by Narla et al. (60), algorithms are more likely to classify images with rulers as malignant because images with rulers were more likely to be malignant. Fifth, the "black box" nature of CNN makes it impossible to interpret how and why CNN arrived

FIGURE 6 | Accuracy of skin tumor classification by our CNN classifier. **(A)** Result of 14-class classification by dermatology trainees, board-certified dermatologists, and CNN classifier. Adapted from Fujisawa et al. (39). **(B)** Result of 2-class classification (benign or malignant). In both classification level, the accuracy of CNN surpassed board-certified dermatologists.

at its output. As an example, Navarrete-Dochent et al. (61) reported that the output of their CNN was affected by the size, rotation, or color tone of images. A similar phenomenon was observed in our system; the output of our CNN was affected by changing the size parameters of the tumor (data not shown). To improve the robustness of the CNN classifier, establishment of an open-access, standardized, large skin tumor image dataset, which includes both rare tumors/subtypes and all ethnicities, is mandatory. Moreover, a robust, standardized measurement method for evaluation and comparison of systems should be established.

FUTURE PERSPECTIVE

AI classifiers for the image classification field have been dramatically improved and made more popular by the introduction of CNN. Many strategies, such as creating ensembles of multiple models (62, 63) or using additional information other than image labels, to improve the accuracy of the classifier outside of increasing the number of images for training have been reported (64). Some studies have reported that CNN algorithms have already surpassed the classification efficacy of dermatologists and, in the near future, AI classifiers may gain sufficient sensitivity and specificity to bear the screening burden for detecting malignant skin tumors. Therefore, some physicians may consider AI as a potential threat, but we believe it to be no more than a diagnostic assistance system due to many limitations detailed in previous studies and difficulty in performance comparisons within published results. Besides such limitations, AI classifiers still have important role in assisting

non-dermatologist physicians, since most skin cancer patients will consult them before being transferred to dermatologists. Early detection and treatment are both still essential in the management of melanoma and, therefore, an efficient AI classifier would help to "detect" patients in the early stage of the disease. Further research is thus required to both improve classification efficacy and develop independent evaluation methodologies to accurately measure system efficacy. Moreover, integration of other medical information such as vital signs, routine blood testing, or even omics data, may give us new insights into the biology or pathology of the disease.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

REFERENCES

1. *United States Cancer Statistics: Data Visualiztions*. Available online at: https://gis.cdc.gov/Cancer/USCS/DataViz.html (accessed May 28, 2019).

2. Gruber SB, Armstrong BK. Cutaneous and ocular melanoma. In: Schottenfeld D, Fraumeni JF, editors. *Cancer Epidemiology and Prevention*. New York, NY: Oxford University Press (2006). p. 1196–229. doi: 10.1093/acprof:oso/9780195149616.003.0063

3. Jemal A, Saraiya M, Patel P, Cherala SS, Barnholtz-Sloan J, Kim J, et al. Recent trends in cutaneous melanoma incidence and death rates in the United States, 1992-2006. *J Am Acad Dermatol*. (2011) 65:S17-25 e11-13. doi: 10.1016/j.jaad.2011.04.032

4. Jemal A, Simard EP, Dorell C. Annual Report to the Nation on the Status of Cancer, 1975-2009, featuring the burden and trends in human papillomavirus(HPV)-associated cancers and HPV vaccination coverage levels. *J Natl Cancer Inst*. (2013) 105:175–201. doi: 10.1093/jnci/djt083

5. Rogers HW, Weinstock MA, Feldman SR, Coldiron BM. Incidence estimate of nonmelanoma skin cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. *JAMA Dermatol*. (2015) 151:1081–6. doi: 10.1001/jamadermatol.2015.1187

6. Services UDoHaH. *The Surgeon General's Call to Action to Prevent Skin Cancer*. US Department of Health and Human Services OotSG. Washington, DC (2014).

7. Karagas MR, Weinstock MA, Nelson HH. Keratinocyte carcinomas (basal and squamous cell carcinomas of the skin. In: Schottenfeld D, Fraumeni JF, editors. *Cancer Epidemiology and Prevention*. New York, NY: Oxford University Press (2006). p. 1230–50. doi: 10.1093/acprof:oso/9780195149616.003.0064

8. Deinlein T, Richtig G, Schwab C, Scarfi F, Arzberger E, Wolf I, et al. The use of dermatoscopy in diagnosis and therapy of nonmelanocytic skin cancer. *J Dtsch Dermatol Ges*. (2016) 14:144–51. doi: 10.1111/ddg.12903

9. Verhoeven EW, Kraaimaat FW, van Weel C, van de Kerkhof PC, Duller P, van der Valk PG, et al. Skin diseases in family medicine: prevalence and health care use. *Ann Fam Med*. (2008) 6:349–54. doi: 10.1370/afm.861

10. Julian CG. Dermatology in general practice. *Br J Dermatol*. (1999) 141:518–20. doi: 10.1046/j.1365-2133.1999.03048.x

11. Kerr OA, Tidman MJ, Walker JJ, Aldridge RD, Benton EC. The profile of dermatological problems in primary care. *Clin Exp Dermatol*. (2010) 35:380–3. doi: 10.1111/j.1365-2230.2009.03586.x

12. Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice Hall (2009).

13. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med*. (2018) 284:603–19. doi: 10.1111/joim.12822

14. Masood A, Al-Jumaily AA. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *Int J Biomed Imaging*. (2013) 2013:323268. doi: 10.1155/2013/323268

15. Mirzaalian H, Lee TK, Hamarneh G. Hair enhancement in dermoscopic images using dual-channel quaternion tubularness filters and MRF-based multilabel optimization. *IEEE Trans Image Process*. (2014) 23:5486–96. doi: 10.1109/TIP.2014.2362054

16. Abbas Q, Garcia IF, Emre Celebi M, Ahmad W, Mushtaq Q. Unified approach for lesion border detection based on mixture modeling and local entropy thresholding. *Skin Res Technol*. (2013) 19:314–9. doi: 10.1111/srt.12047

17. Chang WY, Huang A, Chen YC, Lin CW, Tsai J, Yang CK, et al. The feasibility of using manual segmentation in a multifeature computer-aided diagnosis system for classification of skin lesions: a retrospective comparative study. *BMJ Open*. (2015) 5:e007823. doi: 10.1136/bmjopen-2015-007823

18. Euijoon A, Lei B, Youn Hyun J, Jinman K, Changyang L, Fulham M, et al. Automated saliency-based lesion segmentation in dermoscopic images. *Conf Proc IEEE Eng Med Biol Soc*. (2015) 2015:3009–12. doi: 10.1109/EMBC.2015.7319025

19. Kasmi R, Mokrani K, Rader RK, Cole JG, Stoecker WV. Biologically inspired skin lesion segmentation using a geodesic active contour technique. *Skin Res Technol*. (2016) 22:208–22. doi: 10.1111/srt.12252

20. Aitken JF, Pfitzner J, Battistutta D, O'Rourke PK, Green AC, Martin NG. Reliability of computer image analysis of pigmented skin lesions of Australian adolescents. *Cancer*. (1996) 78:252–7. doi: 10.1002/(SICI)1097-0142(19960715)78:2<252::AID-CNCR10>3.0.CO;2-V

21. Cascinelli N, Ferrario M, Bufalino R, Zurrida S, Galimberti V, Mascheroni L, et al. Results obtained by using a computerized image analysis system designed as an aid to diagnosis of cutaneous melanoma. *Melanoma Res*. (1992) 2:163–70. doi: 10.1097/00008390-199209000-00004

22. Marill T, Green D. On the effectiveness of receptors in recognition systems. *IEEE Trans Inform Theory*. (1963) 9:11–7. doi: 10.1109/TIT.1963.1057810

23. Whitney AW. A direct method of nonparametric measurement selection. *IEEE Trans Comput*. (1971) C-20:1100–3. doi: 10.1109/T-C.1971.223410

24. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. *Pattern Recogn Lett*. (1994) 15:1119–25. doi: 10.1016/0167-8655(94)90127-9

25. Byvatov E, Schneider G. Support vector machine applications in bioinformatics. *Appl Bioinformatics*. (2003) 2:67–77.

26. Krzywinski M, Altman N. Classification and regression trees. *Nat Methods*. (2017) 14:757. doi: 10.1038/nmeth.4370

27. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med*. (1984) 3:143–52. doi: 10.1002/sim.4780030207

28. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press (1996). doi: 10.1017/CBO9780511812651

29. Ferris LK, Harkes JA, Gilbert B, Winger DG, Golubets K, Akilov O, et al. Computer-aided classification of melanocytic lesions using dermoscopic images. *J Am Acad Dermatol*. (2015) 73:769–76. doi: 10.1016/j.jaad.2015.07.028

30. Chang WY, Huang A, Yang CY, Lee CH, Chen YC, Wu TY, et al. Computer-aided diagnosis of skin lesions using conventional digital photography: a reliability and feasibility study. *PLoS ONE*. (2013) 8:e76212. doi: 10.1371/journal.pone.0076212

31. Bhattacharya A, Young A, Wong A, Stalling S, Wei M, Hadley D. Precision diagnosis of melanoma and other skin lesions from digital images. *AMIA Jt Summits Transl Sci Proc*. (2017) 2017:220–6.

32. Russakovsky O, Deng J, Su H. ImageNet large scale visual recognition challenge. *Int J Comput Vision*. (2015) 115:211–52. doi: 10.1007/s11263-015-0816-y

33. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: *25th International Conference on Neural Information Processing Systems*. New York, NY (2012). p. 1097–105. doi: 10.1145/3065386

34. Szegedy C, Liu W, Jia Y. Going deeper with convolutions. In: *IEEE Conference on Computer Vision and Pattern Recognition*. New York, NY: IEEE (2014). doi: 10.1109/CVPR.2015.7298594

35. Arribas JI, Cid-Sueiro J. A model selection algorithm for a posteriori probability estimation with neural networks. *IEEE Trans Neural Netw*. (2005) 16:799–809. doi: 10.1109/TNN.2005.849826

36. Yap MH, Pons G, Marti J, Ganau S, Sentis M, Zwiggelaar R, et al. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE J Biomed Health Inform*. (2018) 22: 1218–26. doi: 10.1109/JBHI.2017.2731873

37. Zeiler MD, Fergus R. *Visualizing and Understanding Convolutional Networks 2013*. Available online at: *arXiv[Preprint].arXiv:1311.2901* (accessed 25 July 2019).

38. Cheng PM, Malhi HS. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *J Digit Imaging*. (2017) 30:234–43. doi: 10.1007/s10278-016-9929-2

39. 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. doi: 10.1111/bjd.16924

40. 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. doi: 10.1038/nature21056

41. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. *Sci Rep*. (2019) 9:6381. doi: 10.1038/s41598-019-42294-8

42. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks 2017. *arXiv[Preprint].arXiv:1709.01507*.

43. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. (2016) 3:9. doi: 10.1186/s40537-016-0043-6

44. Steiner A, Pehamberger H, Wolff K. *In vivo* epiluminescence microscopy of pigmented skin lesions. II. Diagnosis of small pigmented skin lesions and early detection of malignant melanoma. *J Am Acad Dermatol*. (1987) 17:584–91. doi: 10.1016/S0190-9622(87)70240-0

45. Piccolo D, Ferrari A, Peris K, Diadone R, Ruggeri B, Chimenti S. Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis of 341 pigmented skin lesions: a comparative study. *Br J Dermatol*. (2002) 147:481–6. doi: 10.1046/j.1365-2133.2002.04978.x

46. Nachbar F, Stolz W, Merkle T, Cognetta AB, Vogt T, Landthaler M, et al. The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions. *J Am Acad Dermatol*. (1994) 30:551–9. doi: 10.1016/S0190-9622(94)70061-3

47. Ahnlide I, Bjellerup M, Nilsson F, Nielsen K. Validity of ABCD rule of dermoscopy in clinical practice. *Acta Derm Venereol*. (2016) 96:367–72. doi: 10.2340/00015555-2239

48. Johr RH. Dermoscopy: alternative melanocytic algorithms-the ABCD rule of dermatoscopy, Menzies scoring method, and 7-point checklist. *Clin Dermatol*. (2002) 20:240–7. doi: 10.1016/S0738-081X(02)00236-5

49. Rao BK, Ahn CS. Dermatoscopy for melanoma and pigmented lesions. *Dermatol Clin*. (2012) 30:413–34. doi: 10.1016/j.det.2012.04.005

50. Argenziano G, Soyer HP. Dermoscopy of pigmented skin lesions–a valuable tool for early diagnosis of melanoma. *Lancet Oncol*. (2001) 2:443–9. doi: 10.1016/S1470-2045(00)00422-8

51. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol*. (2002) 3:159–65. doi: 10.1016/S1470-2045(02)00679-4

52. Menzies SW, Bischof L, Talbot H, Gutenev A, Avramidis M, Wong L, et al. The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. *Arch Dermatol*. (2005) 141:1388–96. doi: 10.1001/archderm.141.11.1388

53. Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/ artificial intelligence for the diagnosis of melanoma. *Br J Dermatol*. (2009) 161:591–604. doi: 10.1111/j.1365-2133.2009.09093.x

54. Haenssle HA, Fink C, Schneiderbauer R. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. (2018). doi: 10.1093/annonc/mdy520

55. 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*. (2019) 20:938–47. doi: 10.1016/S1470-2045(19)30333-X

56. Han SS, Lim W, Kim MS, Park I, Park GH, Chang SE. Interpretation of the outputs of a deep learning model trained with a skin cancer dataset. *J Invest Dermatol*. (2018) 138:2275–7. doi: 10.1016/j.jid.2018.05.014

57. 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*. (2018) 138:1529–38. doi: 10.1016/j.jid.2018.01.028

58. 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*. (2019) 111:148–54. doi: 10.1016/j.ejca.2019.02.005

59. Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. *Exp Dermatol*. (2018) 27:1261–7. doi: 10.1111/exd.13777

60. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. *J Invest Dermatol*. (2018) 138:2108–10. doi: 10.1016/j.jid.2018.06.175

61. 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. doi: 10.1016/j.jid.2018.04.040

62. Perez F, Avila S, Valle E. Solo or ensemble? Choosing a CNN architecture for melanoma classification. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Long Beach, CA: IEEE Xplore (2019).

63. Mohanraj V, Sibi Chakkaravarthy S, Vaidehi V. Ensemble of convolutional neural networks for face recognition. In: Kalita J, Balas V, Borah S, editors. *Recent Developments in Machine Learning and Data Analytics. Advances in Intelligent Systems and Computing*. Singapore: Springer (2019). doi: 10.1007/978-981-13-1280-9_43

64. Codella N, Lin C, Halpern A, Hind M, Feris R, Smith J. Collaborative human-AI (CHAI): evidence-based interpretable melanoma classification in dermoscopic images. In: *Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI 2018*. (2018). doi: 10.1007/978-3-030-02628-8_11

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.