Research Techniques Made Simple: Deep Learning for the Classification of Dermatological Images

Marta Cullell-Dalmau[1], Marta Otero-Viñas[2,3] and Carlo Manzo[1]

Deep learning is a branch of artificial intelligence that uses computational networks inspired by the human brain to extract patterns from raw data. Development and application of deep learning methods for image analysis, including classification, segmentation, and restoration, have accelerated in the last decade. These tools have been progressively incorporated into several research fields, opening new avenues in the analysis of biomedical imaging. Recently, the application of deep learning to dermatological images has shown great potential. Deep learning algorithms have shown performance comparable with humans in classifying skin lesion images into different skin cancer categories. The potential relevance of deep learning to the clinical realm created the need for researchers in disciplines other than computer science to understand its fundamentals. In this paper, we introduce the basics of a deep learning architecture for image classification, the convolutional neural network, in a manner accessible to nonexperts. We explain its fundamental operation, the convolution, and describe the metrics for the evaluation of its performance. These concepts are important to interpret and evaluate scientific publications involving these tools. We also present examples of recent applications for dermatology. We further discuss the capabilities and limitations of these artificial intelligence-based methods.

Journal of Investigative Dermatology (2020) **140,** 507–514; doi:10.1016/j.jid.2019.12.029

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, AND DEEP LEARNING

Artificial intelligence (AI) describes a branch of computer science that uses machines to simulate cognitive functions of the human mind, such as learning or reasoning (Figure 1). An increasing number of systems based on AI, such as voice-powered assistants like Alexa and Siri, are progressively affecting human habits. Self-driving cars, speech recognition, and machine vision promise to broadly improve human lives, with applications to business, education, and healthcare.

Subcategories of AI include machine learning (ML) and deep learning (DL, Figure 1). ML is based on the acquisition of knowledge from data and does not provide specific rules for a given task; the machine undergoes a learning process based on examples and optimizes its performance on a specific assignment. ML has been successfully applied to several tasks, including classifying gene expression patterns associated with diseases, predicting protein structures from genetic sequences, or designing chemical scaffolds in drug discovery (Marx, 2019). Generally speaking, DL is one of the several computing systems for ML inspired by the biological neural networks that constitute the human brain. DL utilizes artificial neural networks (ANNs), which attempt to mimic how the brain works, especially the connections between neurons. An ANN is formed by a collection of nodes (also called artificial neurons) arranged in layers and connected to transmit signals (Figure 2). Typically, each signal consists of a number, and the output of each node is a nonlinear function of the sum of the inputs. Nodes and connections are characterized by weights that are adjusted through the learning process to increase or decrease the strength of a given signal. The aggregate signal of an artificial node may pass through an activation function, such as transmitting only signals above a threshold (Figure 2a). An ANN may have a single or multiple hidden layers between the input and the output. The number of hidden layers and the number of nodes in each layer constitute the variables controlling the architecture of the network, called hyperparameters. ANNs with several hidden layers are generally referred to as deep neural networks, thus leading to the use of the term deep learning (Figure 2b). However, there is no clear consensus on the minimum number of layers for a network to be qualified as deep. One of the first deep ANNs had only three hidden layers (Hinton et al., 2006). A high number of layers makes DL more capable than traditional ML of modeling complex data. Moreover, DL can automatically discover the features needed to accomplish its task, whereas ML requires being programmed with the criteria defining such features. However,

[1]*QuBI lab, Faculty of Sciences and Technology, University of Vic – Central University of Catalonia, Vic, Spain;* [2]*Tissue Repair and Regeneration Laboratory, Faculty of Sciences and Technology, University of Vic – Central University of Catalonia, Vic, Spain; and* [3]*Faculty of Medicine, University of Vic – Central University of Catalonia, Vic, Spain*

Correspondence: Carlo Manzo, Faculty of Sciences and Technology, University of Vic – Central University of Catalonia, C. de la Laura, 13 - 08500, Vic, Spain.E-mail: carlo.manzo@uvic.cat

Glossary of terms can be found at the end of this paper.

Abbreviations: AI, artificial intelligence; ANN, artificial neural network; AUC, area under the curve; CNN, convolutional neural network; DL, deep learning; ISIC, International Skin Imaging Collaboration; ML, machine learning; ROC, receiver operating characteristic

SUMMARY POINTS

- Inspired by the visual cortex mechanism, convolutional neural networks exploit the information contained in image datasets to automatically learn features and patterns not always identified by humans.
- Deep learning has demonstrated the capability of achieving highly accurate classification of images of skin lesions associated with cancer and other dermatological conditions.
- Deep learning might be a formidable tool to potentially assist dermatologists in their diagnostic decisions.
- Important limitations to the extension of deep learning methods to care practice include the lack of clarity of the automated decision-making process, inherent to convolutional neural networks and concerns about its accuracy, related to the use of not fully representative training datasets or nonstandardized images.

Advantages

- Automated classification of images of skin lesions associated with different diseases with high accuracy.
- Short execution time after training.
- Useful to support clinicians in diagnosis
- Cost saving by reducing unnecessary biopsies or instrumental analysis.

Limitations

- Need for large training datasets including images from different conditions, ethnicities, and settings.
- Need for standardized images associated with precise clinical metadata.
- Obscure decision-making process for classification.
- Limited accuracy and generalizability when trained on datasets with underrepresented conditions.

as a consequence of the higher number of weights to be determined, the training of DL networks requires large quantities of data.

CONVOLUTIONAL NEURAL NETWORK

Study and application of DL has rapidly accelerated in academic research, business, and popular interest. These advances have been based mainly on the use of the convolutional neural network (CNN), an algorithmic architecture inspired by the human visual cortex (Schmidhuber, 2015). Although CNNs were invented in the 1980s (Fukushima, 1980), it was not until the early 2010s that

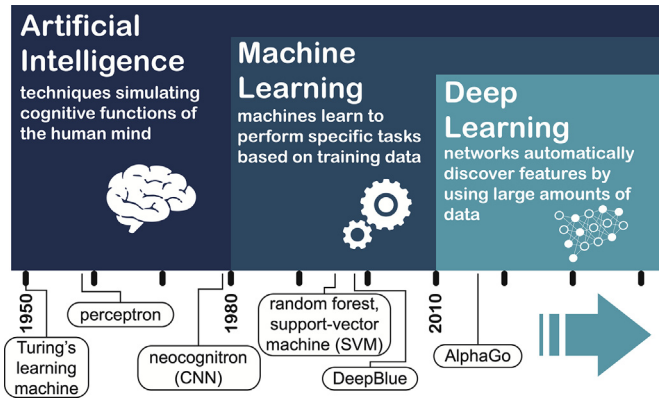

Figure 1. The evolution of artificial intelligence, machine learning, and deep learning. Schematic representation of the timeline and relationship between the three fields, together with a few representative key milestones. CNN, convolutional neural network; SVM, support-vector machine.

massive amounts of labeled data became available for training (Wehner et al., 2017). The growth of computer power deriving from graphics-processing units have fueled massive application of CNNs, in particular as a tool for image classification. As an example of their versatility and power, a CNN constitutes the core of AlphaGo, the computer system that defeated the world's best human player at the game Go (Silver et al., 2016).

As implied by their name, CNNs are mainly based on convolutional layers. Convolution is a mathematical operation between two mathematical functions, which consists in taking element-wise multiplications followed by a sum while shifting one function along the other. CNNs are very well-suited to work with images because of their similarity with

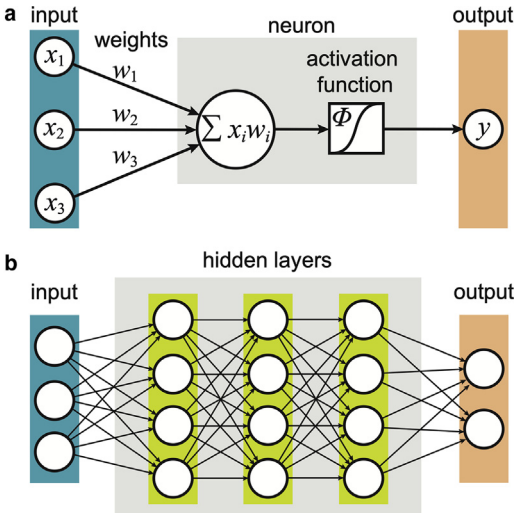

Figure 2. Artificial neurons and neural networks. (**a**) Structure of a node or artificial neuron. The neuron receives inputs from one or more sources, multiplies each of these inputs by a weight, and adds the resulting products. The resulting sum is passed to an activation function and it provides a single output. (**b**) Schematic representation of a basic fully connected network. For illustrative purpose, we show a simple network composed of an input layer, three hidden layers, and an output layer. Each hidden layer is composed of four nodes. The training process creates a model by assigning values to all the weights of the network.

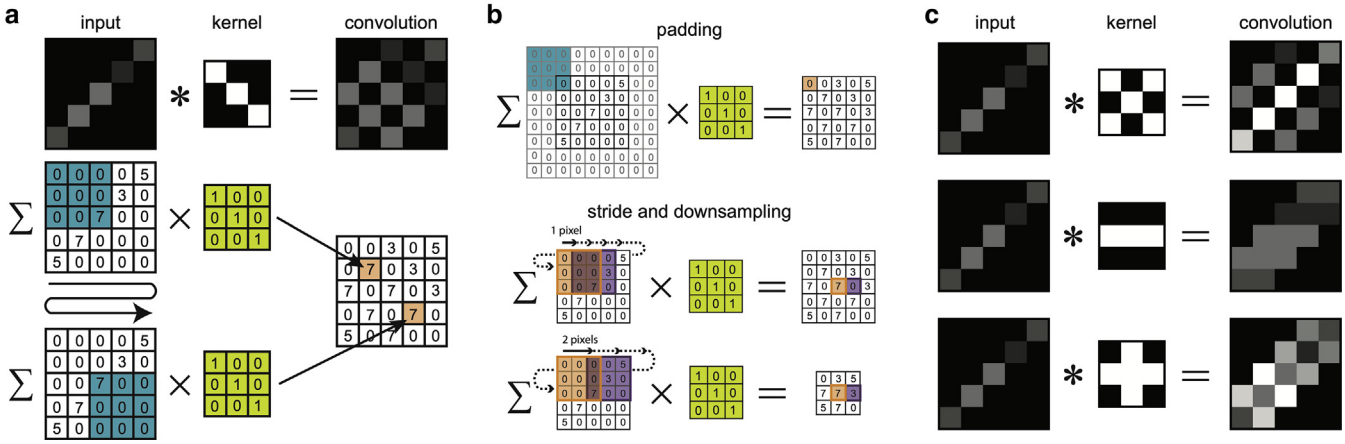

Figure 3. The basics of a CNN: the convolution operation. (**a**) Example of the convolution (*) of a 5 × 5 pixels^2 grayscale image with a 3 × 3 pixels^2 kernel. Grayscale images correspond to numeric matrices, where each pixel is associated with a numeric value. The convolutional feature is obtained by shifting the kernel over the image. At each position, the value of a pixel of the convolutional feature is obtained by multiplying each pixel of the input image by the corresponding pixel of the kernel and then taking the sum. (**b**) Symmetrically padding the input with zeros allows the kernel to operate at the edges of the image and thus preserve the size. Downsampled images can be obtained by changing the stride, the step length at which the kernel is shifted along the input. (**c**) Examples of different features obtained by applying different convolutional kernels to the same input image. CNN, convolutional neural network.

the animal visual cortex, which is composed by neurons that individually respond to small regions of the visual field. An image can be simply viewed as a collection of planes (corresponding to different colors, e.g., 3 for RGB or 1 for grayscale), where each plane is a two-dimensional matrix of numbers (the pixel values).

The convolution of an image plane implies the use of a second matrix, called a kernel, which is shifted along the first one. At each shift position, every pixel of the region of the input image overlapping with the kernel is multiplied by the corresponding pixel of the kernel. The sum of these products produces the value of a pixel of the convolutional feature (Figure 3a). Thus, the pixels of the kernel act like the weights of an artificial neuron over an input corresponding to a region of the input image. The kernel size defines the receptive field of the neuron, that is, the region of the input that is codified into a single value of the output. To make it possible for the kernel to operate at the edges of the input image and preserve the size, convolutional layers generally use zero-padding, the insertion of zero elements around the input image (Figure 3b). Moreover, the convolved image can be obtained by shifting the kernel in steps of one or more pixels. The length of these steps is called stride and, if larger than one, provides an output with smaller lateral size with respect to the original image (Figure 3b). In this way, the convolution can allow for the downsampling of the image while retaining information contained in adjacent pixels.

An important characteristic of convolution is that it can perform different operations on the original image by changing the kernel (Figure 3c). Examples of these operations include blurring, sharpening, denoising, and edge detection. Therefore, a clever combination of randomly selected kernels can lead to the refinement of the computer vision model and, thus, lead to the discovery of new properties.

CNN workflow and model evaluation

To better understand how a CNN works, we will discuss a schematic example from dermatology. Although several types of algorithms have been developed, because of space limitations, we will focus on a supervised learning algorithm for skin lesion classification (Figure 4a). The task of the algorithm is to determine from a digital photograph (input) whether a skin lesion is associated with a malignant cancer or is a benign lesion (output). Because the possible outputs are limited to a finite set of values (only two in this case), this is a (binary) classification problem.

In a typical CNN architecture for classification, the input image is progressively downsampled while increasing the number of kernels and thus obtaining more convolutional features. The last layers have the role of transforming the feature map into a vector, the values of which represent the probability that the image belongs to each class (Figure 4a). In addition to convolutional layers, other layers contribute to perform the mathematical operations necessary to transform the input image and to associate it to the output class. However, their description goes beyond the scope of this article.

In a supervised approach, the algorithm is trained using a labeled dataset, a set of images for which the gold standard output label has been obtained with alternative methods, such as a biopsy. As further detailed in Torres and Judson-Torres (2019), the data are usually split in the following three cohorts: the training set, which is used to determine the weights characterizing the model; the validation set, which is used to assess the model performance during training; and the test set, which is used to evaluate how well the model performs on an unknown input. A CNN iteratively updates the kernel weights of its layers in a random fashion to automatically calculate features from the images and combine them to optimize the connection between the input and the output on the training dataset.

Once the training is complete, the test set is used to quantify the model performance. The simplest quality measure is the classification accuracy, which reports the percentage of correct predictions over the total. However, a high accuracy alone does not guarantee the goodness of a model.

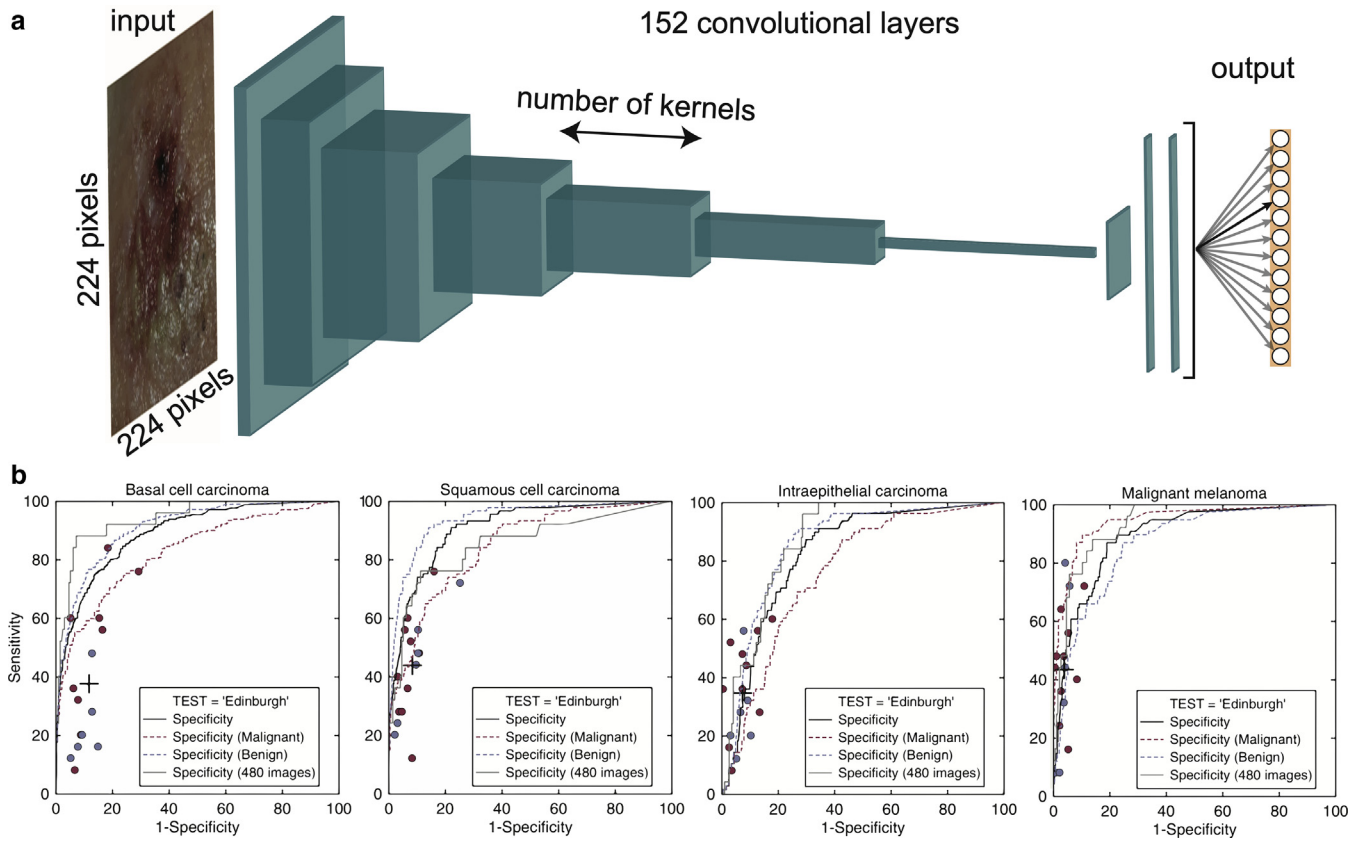

Figure 4. CNN applied to skin cancer classification. (**a**) Scheme of the ResNet-152 CNN used by Han et al. (2018a). Input images with a size of 224 × 224 pixels^2 are analyzed through 152 convolutional layers and classified among 12 different skin diseases. At each block of layers, the CNN progressively downsamples the images while increasing the number of kernels and thus obtaining more convolutional features. The last layers transform the feature map into a vector, the values of which represent the probability that the image belongs to each class. (**b**) ROC curves for the prediction of malignancy in the Edinburgh dataset cases (220 images) reported in Han et al. (2018a). The gray curve corresponds to the results obtained by the ResNet-152 CNN in comparison with 16 dermatologists (red and blue dots). The other curves display the global specificity (black) and the specificity for benign (blue) and malignant conditions (red). CNN, convolutional neural network; ROC, receiver operating characteristic.

For example, a naïve model that always classifies skin lesions as benign will score 96% accuracy on an unbalanced dataset containing 100 images of skin lesions of which only 4 correspond to cancer. The same model will only reach 50% accuracy on a balanced dataset in which the images are equally split between the categories. This ambiguity can be removed by using the confusion matrix, a table reporting the number of correct and incorrect predictions with respect to the actual class. These metrics provide a complete overview of the performance of a model, and its off diagonal elements characterize the level of misclassification.

The typical output of a binary classifier is a numerical value associated with the probability that a given image belongs to the cancer or benign class. A threshold must thus be set to assign an input to one of these two classes based on this probability value. This property allows for the definition of another useful metric for model performance, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve (Figure 4b). The ROC curve is the plot of the sensitivity against the false positive rate (i.e., one minus the specificity) obtained by varying the discrimination threshold used to assign the input to either of the two classes (Figure 4b). It is important to note that the ROC is insensitive to the proportion of the elements contained in each class (Fawcett, 2006). The ROC curve will go from the origin of the axes (0, 0) to (100%, 100%) with a trend that depends on the model behavior. An ROC curve steeply increasing toward high sensitivity at small false positive rates indicates a model that achieves high recall without significantly losing precision. In contrast, an ROC curve increasing with a 45° slope indicates a model with no predictive power. The AUC of the ROC curve can thus be used as a metric to summarize the ROC behavior, because a larger AUC is obtained for models more capable of correctly discriminating between classes.

When extending the problem to a multiclass classification, the confusion matrix further allows for simultaneously visualizing the results of all the classes at a glance. The calculation of the ROC curve becomes a complicated multidimensional problem. A simplification relies on calculating an ROC curve for each class against all the others. However, this approximation removes the insensitivity of the ROC to class imbalance (Fawcett, 2006). An alternative metric for multiclass problems is the top-(n) accuracy, which scores the probability of providing the correct classification within its (n)th choice. In fact, for a given input, a multiclass model will provide probability outputs associated with each class, which will allow ranking of the categories from the most likely (highest output probability) to the least. The top-(1)

accuracy, calculated by taking into account only the prediction associated with the highest output probability, provides the percentage of inputs correctly classified, that is, the standard accuracy. The top-(n) accuracy relaxes this condition by quantifying if the correct class is within the top-(n) outputs provided by the model.

Recent applications to dermatology

Without any prior knowledge about dermatological images, CNNs extract and combine sets of abstract features and automatically generate identifying characteristics (such as a combination of colors, shape, texture, and border geometry) associated with different data categories. In this way, a CNN will learn how to achieve a precise classification of images not included in the training dataset and even find patterns not identified by humans.

In the last years, researchers have started to extensively use DL and CNNs for the analysis of medical images from several disciplines, including dermatology (Esteva et al., 2019; Litjens et al., 2017). Because skin cancer is one of the most common malignancies globally, important efforts have been dedicated to its detection from dermoscopic images only (Codella et al., 2015) or in combination with regular photographic images (Esteva et al., 2017).

To support research and development of methods for automated diagnosis of melanoma, the International Skin Imaging Collaboration (ISIC) has developed a repository of dermoscopic images and it yearly organizes a challenge for the analysis of images of skin lesions (Codella et al., 2018; Marchetti et al., 2018; Tschandl et al., 2019). All the teams taking part in the ISBI melanoma detection challenge in 2016 used DL methods. In 2017, approaches combining DL with additional data led to the highest performance in classification tasks. DL is rapidly becoming the method of choice for image analysis, as testified by the increasing number of publications, especially in the last two years (Brinker et al., 2018). Unquestionably, a milestone was set by the work published in Nature by Esteva et al. (2017), in which a standard CNN architecture (Google's Inception v3) was trained on both dermoscopic and standard photographic images using a dataset of over 100,000 images. The authors proved that the CNN performed similarly to tested experts in classifying malignant versus benign lesions of both epidermal and melanocytic origin. Several other studies have been devoted to the same topic by using other CNN architectures (Fujisawa et al., 2019; Haenssle et al., 2018; Han et al., 2018a). As an example, Figure 4 depicts the architecture of the ResNet-152 CNN used by Han et al. (2018a) and some of the corresponding results. All of these works have reported the equivalence between computer and human diagnosis. Besides skin cancer detection, DL is also being successfully applied to other areas of dermatology, such as the monitoring of wound healing (Shenoy et al., 2018), the classification of ulcers (Goyal et al., 2018), and onychomycosis (Han et al., 2018b).

In addition to classification tasks, DL-based models for the segmentation of skin lesions and ulcers have also been successfully developed (Yap et al., 2019). In particular, these methods have been shown to provide an accurate wound area quantification (Lu et al., 2017; Wang et al., 2015) and promising results on image-based identification of distinct tissues within dermatological wounds (Blanco et al., 2020).

LIMITATIONS AND CHALLENGES

Advances in DL have been accompanied by contrasting reactions. Enthusiastic claims about the outperformance of human diagnosis have been dampened by doubts and criticisms about DL being nothing but an overhyped black box. As always, the truth seems to lie somewhere in between. DL has undoubtedly achieved notable accomplishments in very specific tasks and fields, but it is still far from the realization of a human-equivalent AI.

DL is often considered a black box because its decision-making process is somehow obscured by the thousands of training parameters. In practice, weights and features are often uninterpretable and it is thus difficult for the researchers to fully grasp the working process of a model or the reason why it provides specific performance. The extent to which the inner working of a CNN can be explained in human terms is referred to as explainability. Improving explainability represents a key point for AI to ultimately make decisions on behalf of humans in critical areas, such as in health care. Efforts for gaining insight into why a CNN made a specific decision involve the development of methods to visualize what a CNN sees, such as saliency maps that simplify CNN feature maps into a more meaningful representation.

Because DL approaches are data-driven, their principal limitations often come from the data themselves. A usual criticism concerns the need for large labeled datasets. However, the development of transfer learning has relaxed this requirement by introducing the ability to reuse a model developed for a task and trained on a large dataset as the starting point of a new model with a different task.

Beyond the role of the amount of data, the work of Han et al. (2018a) triggered an interesting discussion about the composition of the training dataset. A letter to the editor of the *Journal of Investigative Dermatology* (Navarrete-Dechent et al., 2018) raised concerns about the generalizability of automated diagnosis when the training dataset presents limitations in the spectrum of human populations and/or clinical presentation, as well as variability in image acquisition settings and limited clinical metadata. Indeed, the underrepresentation of clinical or demographic categories is a common and often inherent problem in healthcare-related data, and it might limit the generalizability of a model.

The inclusion of metadata containing sociodemographic information about the patient (sex, skin type, race, and age) is thus necessary to verify the presence of biases related to imbalance or underrepresentation (Navarrete-Dechent et al., 2018). When possible, the obvious solution to this problem is to broaden the dataset by including images and data of patients from less represented groups. As an alternative, the robustness of a model requires further validation, such as through prospective studies.

An inherent weakness of many of the DL models applied so far to dermatology resides in the lack of a "none of the above"

output. If presented with an image not corresponding to any of the training classes, a model will force it into one of the other categories. In this case, to deal with this issue and prevent misclassification, it is thus necessary to use an approach enabling open set recognition.

The lack of standardization of dermatological images represents a strong limitation that affects the development of the research in this field and undermines its integrity and reproducibility. The variability of dermatological images is due to several causes, such as the type of device used to acquire the images, the image acquisition conditions, the amount and type of metadata, and the lack of a standard terminology. Establishing common criteria for data collection and management is fundamental for the creation of large datasets and their sharing between systems and users. Moreover, the lack of standardization, together with the opacity of the CNN inner process, poses a problem for the operation of classifiers. For example, if images of lesions associated with a specific pathology are generally taken at a high resolution, a CNN might learn to detect the high resolution instead of discriminating the right diagnosis.

An effort toward the establishment of standardized conditions is being carried out by the ISIC to ensure image quality, privacy, and interoperability. The project includes the creation of a public archive of images (https://isic-archive.com) to permit independent assessment of the performance of any software. According to ISIC guidelines, images should comply to standards belonging to three categories, technology, technique, and terminology. Furthermore, the presence of detailed metadata including device characteristics, photograph settings, and information about both the patient and the skin lesion is of paramount importance to take full advantage of the information contained in the images. However, the large number of images needed for training further imposes the development of a quality test to automatically assess whether an image respects such quality standards.

Besides image standardization, another strategy might involve the use of an algorithm to intrinsically take this variability into account by introducing an ad hoc augmentation procedure capable of artificially creating variations of brightness, camera angle, body geometry, and skin background, or even introducing rulers, as observed in actual images. Variability sources associated with technical and geometrical parameters might either be measured separately or estimated from the image itself and thus corrected or accounted for by an image preprocessing step. An effective contribution in this sense might come from other DL architectures that are able to infer information such as depth or shape from regular images.

The importance of clinical metadata deserves to be further stressed, because it has also been shown that combining lesion images with sociodemographic data (age and sex), clinical variables (location of the lesion), and close-up images improved the performance of a classifier (Haenssle et al., 2018).

In conclusion, DL and CNN have demonstrated the capability of achieving highly accurate diagnoses in the classification of skin cancer and other dermatological conditions. DL constitutes a formidable tool to potentially assist dermatologists in their clinical decisions. The computer science and dermatology communities are fruitfully collaborating to develop novel approaches toward dermatologic diagnosis. However, the use of DL in healthcare practices still requires further substantiation by data and prospective studies to obtain the acceptance of patients and physicians. For this

MULTIPLE CHOICE QUESTIONS

1. Which of the following statements about artificial intelligence is FALSE?
   - A. It is a branch of computer science.
   - B. It is a synonym of deep learning.
   - C. It includes machine learning and deep learning as subcategories.
   - D. It uses machines for simulating cognitive functions of the brain.

2. The advantages of convolutional neural networks do NOT include:
   - A. Automated image classification with high accuracy.
   - B. Once training is done, it achieves fast classification.
   - C. It combines abstract features to find patterns.
   - D. Fast training by using small labeled databases, publicly available.

3. Which of the following statements about convolutional neural network datasets is TRUE?
   - A. They are usually divided into three groups for training, validation, and test.
   - B. Relatively large datasets are needed.
   - C. It needs to be labeled with the correct output.
   - D. All of the above.

4. Which of the following quantities are usually used to evaluate the performance of a classifier?
   - A. The area under the receiver operating characteristic curve.
   - B. The ratio between sensitivity and specificity.
   - C. The false positive rate at varying thresholds.
   - D. The Jaccard index.

5. Deep learning is often dubbed “black box” because:
   - A. It is commonly used as a flight recorder.
   - B. It is the name of the company that first used this technology.
   - C. Its decision-making process is obscured by the thousands of training parameters.
   - D. It is the color of its shipment case.

See online version of this article for a detailed explanation of correct answers.

reason, a careful risk evaluation should be assessed before making publicly available any research tool without a prospective validation (Narla et al., 2018).

ORCIDs

Marta Cullell-Dalmau: https://orcid.org/0000-0001-5469-0826
Marta Otero-Viñas: https://orcid.org/0000-0003-2718-9977
Carlo Manzo: https://orcid.org/0000-0002-8625-0996

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

CM gratefully acknowledges funding from FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación through the "Ramón y Cajal" program 2015 (Grant No. RYC-2015-17896), and the "Programa Estatal de I+D+i Orientada a los Retos de la Sociedad" (Grant No. BFU2017-85693-R) from the Generalitat de Catalunya (AGAUR Grant No. 2017SGR940). CM also acknowledges the support of NVIDIA Corporation with the donation of the Titan Xp GPU. MO-V gratefully acknowledges funding from the PO FEDER of Catalonia 2014-2020 (project PECT Osona Transformació Social, Ref. 001-P-000382) and the Spanish Ministry of Science, Innovation, and Universities through the Instituto de Salud Carlos III-FEDER program (FIS PI19/01379).

AUTHOR CONTRIBUTIONS

Conceptualization: CM, MC-D; Funding Acquisition: CM, MO-V; Supervision: CM, MO-V; Visualization: CM, MC-D; Writing - Original Draft Preparation: CM, MC-D; Writing - Review and Editing: CM, MO-V.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to this paper. Teaching slides are available as supplementary material.

REFERENCES

Blanco G, Traina AJM, Traina C Jr, Azevedo-Marques PM, Jorge AES, de Oliveira D, et al. A superpixel-driven deep learning approach for the analysis of dermatological wounds. Comput Methods Programs Biomed 2020;183:105079.

Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 2018;20:e11936.

Codella N, Cai J, Abedini M, Garnavi R, Halpern A, Smith JR. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In: Zhou L, Wang L, Wang Q, Shi Y, editors. Machine learning in medical imaging. Lecture notes computer science series. Cham: Springer International Publishing; 2015. p. 118–26.

Codella NCF, 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). In: 2018 IEEE 15th international symposium Biomedicine Imaging (ISBI 2018). New York, NY: IEEE; 2018. p. 168–72.

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

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

Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27:861–74.

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.

Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 1980;36:193–202.

Goyal M, Reeves ND, Davison AK, Rajbhandari S, Spragg J, Yap MH. DFU-Net: convolutional neural networks for diabetic foot ulcer classification. IEEE Trans Emerg Top Comput Intell 2018:1–12.

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.

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

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

Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput 2006;18:1527–54.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.

Lu H, Li B, Zhu J, Li Y, Li Y, Xu X, et al. Wound intensity correction and segmentation with convolutional neural networks. Concurr. Comput Pract Exp 2017;29:e3927.

Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A, et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78:270–7.e1.

Marx V. Machine learning, practically speaking. Nat Methods 2019;16:463–7.

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.

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.

Schmidhuber J. Deep Learning in neural networks: an overview. Neural Netw 2015;61:85–117.

Shenoy VN, Foster E, Aalami L, Majeed B, Aalami O. Deepwound: automated postoperative wound assessment and surgical site surveillance through convolutional neural networks. 2018 IEEE international conference Bioinformatics Biomedicine. New York, NY: IEEE; 2018. p. 1017–21.

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9.

Torres R, Judson-Torres RL. Research techniques made simple: feature selection for biomarker discovery. J Invest Dermatol 2019;139:2068–74.e1.

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.

Wang C, Yan X, Smith M, Kochhar K, Rubin M, Warren SM, et al. A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks. 2015 37th Annual international conference IEEE engineering in medicine and biology society. New York, NY: IEEE; 2015. p. 2415–8.

Wehner MR, Levandoski KA, Kulldorff M, Asgari MM. Research techniques made simple: an introduction to use and analysis of big data in dermatology. J Invest Dermatol 2017;137:e153–8.

Yap MH, Goyal M, Ng J, Oakley A. Skin lesion boundary segmentation with fully automated deep extreme cut methods. In: Gimi B, Krol A, editors. Medical Imaging 2019Biomedical applications in molecular, structural, and functional imaging. SPIE; 2019. p. 24.

Glossary

Term	Description
Activation function	A nonlinear function that controls the magnitude of the output signal of a node
Artificial neural network	A brain-inspired computing system that learns to perform tasks by considering examples
Connection	A link between nodes; it transmits the (modified) output signal of a node as the input of another
Convolution	A mathematical operation consisting of the sum of element-wise products between an image and a kernel while shifting one along the other
Convolutional neural network	A class of artificial neural network using the mathematical operation called convolution; they are inspired by the function of the human visual cortex and are well-suited for image analysis
Hidden layer	A layer positioned between the input and the output layer of a network
Kernel	A matrix, generally small, used to extract features from an image through convolution
Layer	A collection of nodes operating simultaneously in the network sequence of tasks
Learnable parameters	Parameters, like weights and biases, that are adjusted during the training process to improve a model
Node, or artificial neuron	The basic unit of a neural network that performs an operation over one or more input signals to produce an output
Stride	The step length in pixels of the kernel shift along the input image during the convolution
Weight	The numerical value associated with a connection that modifies the value of the incoming signal; weights are adjusted during the learning process to strengthen or inhibit specific signals

DETAILED ANSWERS

1. **Which of the following statements about artificial intelligence is FALSE?**

Answer: B. Artificial intelligence and deep learning are not synonyms. Deep learning is a subcategory of machine learning, which in turn is a subcategory of artificial intelligence.

2. **The advantages of convolutional neural networks do NOT include:**

Answer: D. In general, the training is not a very fast process because it requires optimization over a large amount of labeled data, often difficult to obtain.

3. **Which of the following statements about convolutional neural network datasets is TRUE?**

Answer: D. Large labeled datasets are needed to train the great number of parameters of convolutional neural networks. The data are split into three groups and used for training, validation, and testing the model performance.

4. **Which of the following quantities are usually used to evaluate the performance of a classifier?**

Answer: A. The area under the curve of the receiver operating characteristic measures how good a model is in distinguishing between classes.

5. **Deep learning is often dubbed "black box" because:**

Answer: C. The information about the model is encoded inside the values of the weights and is difficult to interpret.