Artificial intelligence (AI) is a hot topic in many sciences, including medicine. A machine capable of making a medical diagnosis in full autonomy is fascinating and highly attractive. “Many years ago, in the movie *Star Trek*, the spaceship’s doctor diagnosed a pancreatic cancer simply by placing a small device on the belly of the patient” recalls Giuseppe Argenziano, professor of dermatology and head of the Dermatology unit at University of Campania, Naples (Italy). He envisages a similar scenario for the dermatology of tomorrow. “One day, in a remote country, anywhere in the world, a general practitioner will receive a patient, notice that they have a black skin lesion, place a little machine on it and get in 10 seconds the answer if the lesion is benign or malignant. We will get there, eventually. The future is really promising, but we are not there yet”.

AI loves images

In the *Encyclopaedia Britannica*, AI is defined as “the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings”. ‘Machine learning’ (ML), is the branch of AI devoted to the development of algorithms that attempt to simulate the human decision process. Trained with billions and billions of data points, the algorithm learns from examples and can be applied without human supervision of unseen data. The most sophisticated ML models that the scientist working on in these days are generated through ‘deep learning’, using artificial neural networks that can learn extremely complex relationships between features and labels (Rajkomar A, et al. NEJM 2019).

Since digitised images are optimal for computer inputs, and since examining many images is not only a tedious and time-consuming task, but basically unfeasible with very high number of data, it was crucial to think about exploiting AI to automate visual data-based screenings. Many research groups around the world are working to apply this approach to lung cancer screening using CT scans (in most tumours via PET and CT and, in the topic we address here, to skin cancer screening, using pictures of lesions (Thomsen K, et al. J Dermatol Treat 2019).

Machines go to school

A critical point in the development of ML-based predictive models is machine training. Which images should be used? Dermoscopic images or clinical close-ups? Standardised images or unstandardised ones, that are closer to real-life input? This question may not have one single answer. Given the GIGO principle (garbage in, garbage out), which states that the quality of output is determined by the quality of the input, the best way to train an algorithm is not unequivocal. One may think that, if the purpose of the tool is to make diagnostics widely accessible, it should probably be trained with photos taken in real-life settings. Conversely, if the goal is to create a decision aid for clinicians, the use of high quality dermoscopic images would guarantee greater accuracy.

Optimal predictive models have high generalisability. Therefore, results may be hampered when the model fits the training data too closely (overfitting). If the training is too overfitted, the algorithm will have a very high performance with the training set, but a poor to mediocre performance on unseen data (Massi D, et al. Lancet Oncol, 2019).

Variability is another important element to be considered. A system robust enough to handle images coming from different sources needs to be able to handle ‘noisy’ data; otherwise, the algorithm will use features of unstandardised photos to guide decision making. The experience of a group of researchers at Stanford University, reported in a *Journal of Investigative Dermatology* paper (Narla A, et al. J Invest Dermatol, 2018), illustrates this problem well.

> “In our work, we noted that the algorithm appeared more likely to interpret images with rulers as malignant. Why? In our dataset, images with rulers were more likely to be malignant; thus, the algorithm inadvertently ‘learned’ that rulers are malignant” says Justin M. Ko, Director and Chief of the Medical Dermatology of Standford Health Care at Stanford University, California. “These biases in AI models are inherent unless specific attention is paid to address inputs with variability”

Human vs machine

“When you develop a system of automated diagnosis, it is obvious that your benchmark is the human being Argenziano says. Many studies have compared the accuracy of human readers versus machine-learning algorithms for skin lesion classification. Results are amazing: state-of-the-art ML classifiers generally outperformed human experts in the diagnosis of skin lesions (Esteva A, et al, Nature 2017; Brinker TJ, et al. Eur J Canc 2019; Brinker TJ, et al. Eur J Canc 2019; Hekler A, et al. Eur J Canc 2019; Tschandl P, Lancet Oncol. 2019; Tschandl P, et al. JAMA Dermatol 2019). This success needs to be contextualised.

> *“Some of these algorithms perform better than I do, despite the fact that I have been practising for thirty years. However, this is true in the experimental setting. If you submit me an image, I perform worse than the computer. In a real-life setting, where I meet the patient and I am aware that they have hundreds of moles, or moles in a certain area of the body, or have a history of melanoma and so on, my diagnostic ability is probably better than the machine’s one”, Argenziano affirms.*

Tactile experience may also be very helpful in the discrimination of malignant lesions, but this cue is completely missing in a visual-only inspection. Philipp Tschandl, dermatologist and researcher at the Medical University of Vienna, Vienna (Austria), author of seminal papers on ML-based algorithms for skin lesion classification, underlines that “although ML algorithms outperformed human experts in nearly every aspect, higher accuracy in a diagnostic study with digital images does not necessarily mean better clinical performance or patient management. Within studies one has usually a very controlled environment, but real life is different. Examiners may not choose the correct spots, the chosen lesions may be of a different distribution (e.g. skin type), etc. Also, one has more information to come to a decision (e.g. is something new or changed? or did the patient have a melanoma previously?).”

In the end, what really matters it is not the performance of algorithms, but whether the use of technology brings benefits to patients, as Federico Cabitza, associate professor of informatics at the

University of Milano-Bicocca, Milan (Italy) argues in an article on the unintended consequences of machine learning in medicine: “The quality of any machine-learning decision support system and subsequent regulatory decisions about its adoption should not be grounded only in performance metrics, but rather should be subject to proof of clinically important improvements in relevant outcomes compared with usual care, along with the satisfaction of patients and physicians”. (Cabitza F, et al. JAMA 2017)

Who needs it?

When asked what would be the most appropriate use of AI-based skin lesion classification tools, Philipp Tschandl replies: “If accurate and safe enough, maybe pre-screening in telemedicine, or guidance of nonexperts. I do not expect accurate and safe-enough techniques allowing fully autonomous self-screening, self-diagnosis in the visible future.”

Giuseppe Argenziano does not think that these AI tools can totally replace the specialist’ check, at least with the available technology, but they can be helpful for other healthcare workers. “A dermatologist has the expertise to correctly diagnose 99% of the lesions seen in daily practice. I cannot imagine a dermatologist used to examine a patient in few minutes with dermatoscope to shift to this technology – now quite time-consuming – to perform the task. This technology may possibly be useful in 5% of cases, but at that point the dermatologist would probably decide for a biopsy.” He emphasises that it is unlikely that, if in doubt, the specialist would rely only on a machine – at least for now. “However, in the future, these tools will become an additional element to be considered, together with all the other factors analysed by the physician. In the end, the responsibility is up to the physician that makes the diagnosis and decide the treatment. The machine has to be seen as a second opinion,” he says.

AI may, however, be very helpful to physicians with little experience in dermatology, such as general practitioners, says Argenziano. While this may not be relevant to countries like Italy, where all skin cancer screening is carried out by specialists, there are other countries, such as Australia, UK and Switzerland, where it could be relevant, because there are few specialists. “In those countries, general practitioners are actively involved in skin cancer screening and they would take advantage of these new tools.” The rise in skin cancer incidence and shortage of specialists pave the road to the use of automated systems.

This opinion is confirmed by Arieh Gomolin, first author of a recent review on applications of AI in dermatology published on Frontiers in Medicine (Gomolin A. et al. Front Med 2020).

“A proper history followed by a physical examination in a well-lit examining room, while assessing for texture and eliciting specific signs for a given lesion, complemented by additional investigations/imaging or a biopsy is a standard way to establish a diagnosis in dermatology” he says. “Furthermore, it is accepted that while some diagnoses are clinical, others rest solely on histologic findings or a combination of clinical and histologic results correlation. This holistic approach cannot be fully replaced by computer programs and this is felt to be one of the most important barriers to implementing AI. Many patients also want to see and partner with a physician who is vested in helping them and may not be satisfied with isolated computerized tools”.

Smartphones, a curse and a blessing

Smartphone applications (apps) are probably the most questionable facet of the attempt to automate skin lesion screening. Several apps have already been developed with the purpose of enabling individuals to assess and track their skin lesions. Most of them do not incorporate AI technology, but are used for a sort of ‘domestic teledermatology’: the image is sent to a dermatologist, who will

make a recommendation. At present, apps that incorporate machine learning-based algorithms are considered unreliable due to poor diagnostic sensitivity. Not surprisingly, these apps usually explicitly state in their terms and conditions that they do not provide a diagnostic service nor do they intend to replace or substitute visits to healthcare providers (Du-Harpur X, et al. Br J Dermatol 2020; Mar VJ, et al. Ann Oncol 2018).

A systematic review on the accuracy of six different algorithm-based smartphone apps to assess risk of skin cancer in suspicious skin lesions has been recently published in *BMJ* (Freeman K, et l. BMJ 2020).

Interestingly, four out of the six apps are no longer available: two of them (DrMole and Spotmole) simply disappeared from the stores, while another two (MelApp and Mole Detective) were withdrawn from the market after that the American Federal Trade Commission fined the marketers for "deceptively claiming the apps accurately analysed melanoma risk". Two apps (SkinVision and SkinCare), which carry the safety, health and environmental CE certification mark, are still available for downloading. "Our review found poor and variable performance of algorithm-based smartphone apps, which indicates that these apps have not yet shown sufficient promise to recommend their use," write the authors. They also highlight that "the current regulatory process for awarding the CE marking for algorithm-based apps does not provide adequate protection to the public."

The regulatory issue is certainly something that will need to be rapidly addressed by the authorities.

"Overdiagnosis and overtreatment are an often overlooked problem that may be an issue with widespread layperson-use of algorithm-based tools," Philip Tschandl warns. A delay in seeking medical advice can likewise have devastating consequences. In addition, smartphone apps are designed for automated solitary lesion classification. They give an answer, right or wrong, to the inquiry. That is all. Therefore "the tool may reassure a hypothetical patient about the lentigo on her arm, while missing the melanoma on her leg" (Narla A, et al. J. Invest. Dermatol. 2018).

Argenziano, too, is very critical about consumer-facing technology.