Quality Improvement Study

Medicine®

OPEN

Adversarial attack on deep learning-based dermatoscopic image recognition systems

Risk of misdiagnosis due to undetectable image perturbations

Jérôme Allyn, MD[a,b,*], Nicolas Allou, MD[a,b], Charles Vidal, MD[a], Amélie Renou, MD[a], Cyril Ferdynus, PhD[b,c,d]

Abstract

Deep learning algorithms have shown excellent performances in the field of medical image recognition, and practical applications have been made in several medical domains. Little is known about the feasibility and impact of an undetectable adversarial attacks, which can disrupt an algorithm by modifying a single pixel of the image to be interpreted. The aim of the study was to test the feasibility and impact of an adversarial attack on the accuracy of a deep learning-based dermatoscopic image recognition system.

First, the pre-trained convolutional neural network DenseNet-201 was trained to classify images from the training set into 7 categories. Second, an adversarial neural network was trained to generate undetectable perturbations on images from the test set, to classifying all perturbed images as melanocytic nevi. The perturbed images were classified using the model generated in the first step. This study used the HAM-10000 dataset, an open source image database containing 10,015 dermatoscopic images, which was split into a training set and a test set. The accuracy of the generated classification model was evaluated using images from the test set. The accuracy of the model with and without perturbed images was compared. The ability of 2 observers to detect image perturbations was evaluated, and the inter observer agreement was calculated.

The overall accuracy of the classification model dropped from 84% (confidence interval (CI) 95%: 82–86) for unperturbed images to 67% (CI 95%: 65–69) for perturbed images (Mc Nemar test, $P < .0001$). The fooling ratio reached 100% for all categories of skin lesions. Sensitivity and specificity of the combined observers calculated on a random sample of 50 images were 58.3% (CI 95%: 45.9–70.8) and 42.5% (CI 95%: 27.2–57.8), respectively. The kappa agreement coefficient between the 2 observers was negative at -0.22 (CI 95%: −0.49–−0.04).

Adversarial attacks on medical image databases can distort interpretation by image recognition algorithms, are easy to make and undetectable by humans. It seems essential to improve our understanding of deep learning-based image recognition systems and to upgrade their security before putting them to practical and daily use.

Abbreviations: AI = artificial intelligence, CI = confidence interval, GPU = graphics processing unit, RAM = random access memory.

Keywords: adversarial attack, artificial intelligence, deep learning, dermatoscopic lesions, image recognition systems

1. Introduction

Deep learning is the most commonly proposed artificial intelligence technology for improving medical care today. The medical field of application that has received most attention is image recognition, with numerous studies highlighting the excellent performance of deep learning algorithms for the classification and analysis of medical images using eye fundus imaging, X-rays, magnetic resonance imaging, or echocardiography.[1–7] These algorithms have been found to improve the diagnostic performance of imaging techniques by reducing the risk of false negative and false positive results. They have also been proposed to help save time and money by automating routine tasks normally performed by medical doctors. Lastly, smartphone applications have been developed that provide rapid diagnosis of skin lesions (among others) without medical advice. However, despite such promise, deep learning algorithms are difficult and sometimes even impossible to understand—a phenomenon often referred to as the "black box" problem.[1,8,9] Moreover, concerns have been raised about their vulnerability to malicious attacks by adversarial networks.[10] The latter do not target the deep learning algorithm itself but the image to be interpreted, as a minimal image perturbation can alter

Editor: Mihnea-Alexandru Găman.

The authors declare that they have no funding and conflicts of interests.

The datasets generated during and/or analyzed during the current study are publicly available.

[a] *Intensive care unit,* [b] *Clinical Informatic Department,* [c] *Methodological Support Unit, Saint-Denis University Hospital, Saint-Denis, Reunion Island,* [d] *INSERM, CIC 1410, F-97410, Saint-Pierre, France.*

[*] *Correspondence: Dr. Jérôme Allyn, Réanimation Polyvalente, Centre Hospitalier Universitaire La Réunion, Site Félix Guyon, Bellepierre 97405 Saint-Denis cedex, France (e-mail: allyn.jer@gmail.com).*

Copyright © 2020 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial License 4.0 (CCBY-NC), where it is permissible to download, share, remix, transform, and buildup the work provided it is properly cited. The work cannot be used commercially without permission from the journal.

How to cite this article: Allyn J, Allou N, Vidal C, Renou A, Ferdynus C. Adversarial attack on deep learning-based dermatoscopic image recognition systems: risk of misdiagnosis due to undetectable image perturbations. Medicine 2020;99:50(e23568).

Received: 30 June 2020 / Received in final form: 30 October 2020 / Accepted: 3 November 2020

http://dx.doi.org/10.1097/MD.0000000000023568

interpretation by the algorithm.[10,11] Such hacking techniques, known as adversarial neural network, have so far been used mainly in the fight against facial recognition.[12–15] To our knowledge only one study related to the medical field, and its results need to be confirmed.[16] Our hypothesis is that adversarial neural network can corrupt diagnosis in the field of dermatology through causing perturbations of skin lesion images that are undetectable by the human eye.

2. Methods

2.1. Study design

The data used for this work come from a database that obtained the necessary ethical approval (University of Queensland, Protocol-No. 2017001223 and Medical University of Vienna Protocol-No. 1804/2017, the data was anonymous and not identifiable, and the source article does not specify the terms of consent).[17] All methods were carried out in accordance with relevant guidelines and regulations.

The open source image database HAM-10000 was split into a training set and a test set. First, the pre-trained convolutional neural network DenseNet-201 was trained to classify images from the training set into 7 categories.[18] The accuracy of the generated classification model was evaluated using images from the test set. Second, an adversarial neural network was trained to generate undetectable perturbations on images from the test set. The perturbed images were classified using the model generated in the first step, and the accuracy of the model was assessed once again. Third, the accuracy of the model with and without perturbed images was compared. Fourth, the ability of 2 observers to detect image perturbations was evaluated, and interobserver agreement was calculated.

2.2. Dataset collection

HAM-10000, short for "Human Against Machine with 10,000 training images," is a publicly available dataset containing 10,015 dermatoscopic images of 7 pigmented skin lesions, namely melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular skin lesions, and dermatofibroma.[17] Image resolution is 600x450 pixels. All images were collected over a period of 20 years from 2 different sites: The Department of Dermatology at the Medical University of Vienna, Austria, and the skin cancer practice of Cliff Rosendahl in Queensland, Australia.

The HAM-10000 dataset was split into a training set and a test set: 80% of the images were used to train the neural network and the other 20% were used to assess the accuracy of the generated classification model. Images were rescaled to 256x256 pixels and normalized between 0 and 1. Augmentation techniques (rotation, cropping) were used to improve the accuracy of the classification model.

2.3. Classification model

Images from the HAM-10000 dataset were classified using DenseNet-201. This convolutional neural network has been trained on more than ten million images from the Image Net database and has been used in the field of healthcare for the diagnosis of breast abnormality.[19,20] DenseNet-201 is 201 layers deep and can classify images into 1000 object categories (pencil, animals, etc.). We modified the last layer of the DenseNet-201 network to make it output the 7 categories of the HAM-10000 dataset (melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular skin lesions, and dermatofibroma). We then trained the network to classify images from the training set through 100 training steps. Lastly, we assessed the accuracy of the generated model on images from the test set by calculating recall and precision statistics for all categories of skin lesion.

2.4. Adversarial model

Following Poursaeed et al, we generated image-dependent perturbations undetectable by the human eye using a pre-trained adversarial neural network.[21] The latter was composed of multiple convolutional layers (Fig. 1): 3 down-sampling

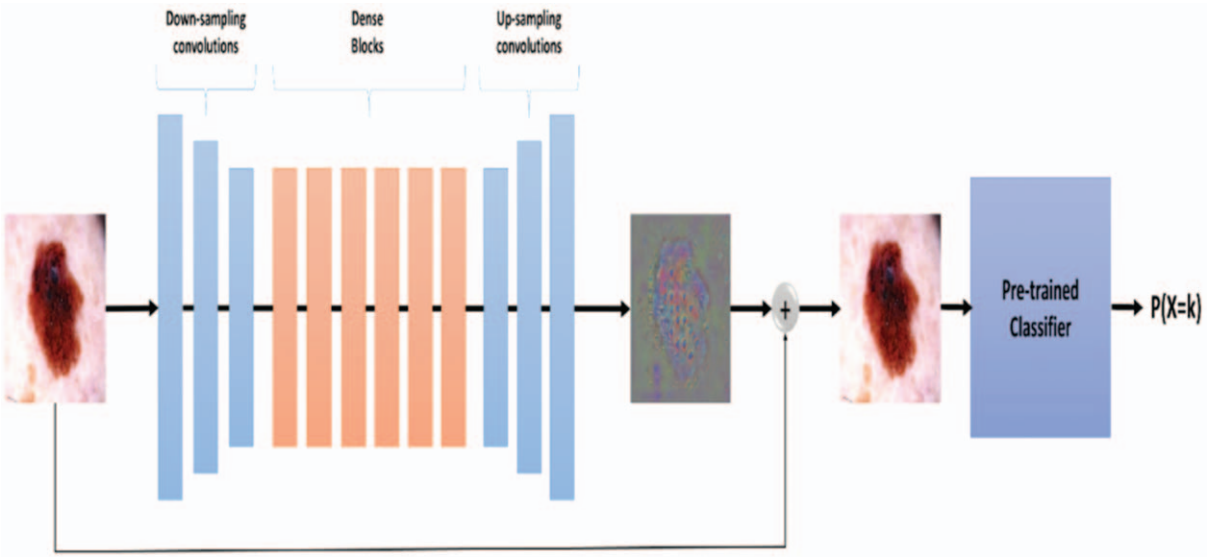

Figure 1. Adversarial neural network architecture (following Poursaeed *et al.*).[21]

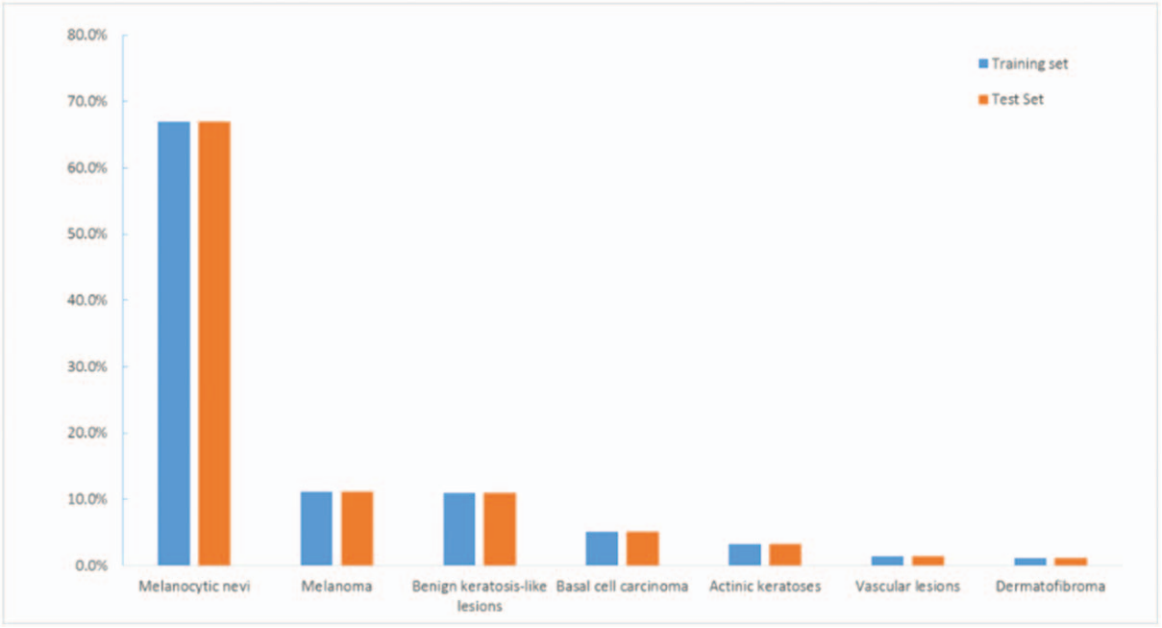

Figure 2. Distribution of the 7 types of skin lesion in the training and test sets.

convolution layers, 6 Dense Blocks layers with residual connections, and 3 up-sampling convolution layers that rescaled generated perturbations to original image size. We added the generated perturbations to images from the test set, while seeking to minimize differences between the original images and the perturbed images. The adversarial network was trained through 200 training steps to fool the classification model into classifying all perturbed images as melanocytic nevi.

2.5. Statistical analysis

We calculated the overall accuracy (i.e., the proportion of well-classified images) of the classification model on images from the test set before and after perturbation, and then compared these images using the McNemar test. A P value under .05 was considered significant. We estimated recall and precision statistics, which are commonly used in the field of machine learning, before and after perturbation. Recall is the proportion of True Positive divided by the sum of True Positive and False Positive (i.e., Predictive Positive Value), while precision is the proportion of True Positive divided by the sum of True Positive and False Negative (i.e., Sensitivity).

We calculated the fooling ratio of the adversarial network overall and for each category of skin lesions by dividing the number of images misclassified as melanocytic nevi (i.e., the target category of the adversarial network) by the total number of images. Variation in overall accuracy before and after perturbation was also evaluated using the McNemar test.

The human ability to recognize a perturbed image was evaluated by asking 2 observers who were blinded to each other to interpret a sample of 50 images randomly chosen from the dataset. Sensitivity and specificity of the combined observers were calculated with their confidence intervals at 95%. Agreement between the 2 observers was assessed by calculating the kappa agreement coefficient with its confidence interval at 95% (CI 95%).

All experiments were coded in Python 3.7 with Tensorflow 2.0 on a personal computer with NVIDIA GeForce 1080 Ti Graphics Processing Unit (GPU) with 12 Gb of Random Access Memory (RAM) (NVIDIA Corp, Santa Clara, CA, USA).[22]

3. Results

We included 8012 images in the training set and 2003 in the test set. Figure 2 shows the distribution of the 7 types of skin lesion for each set. The dataset was unbalanced, with melanocytic nevi representing two-thirds of all images.

The overall accuracy of the classification model dropped from 84% (confidence interval (CI): 95%: 82–86) for unperturbed images to 67% (CI 95%: 65–69) for perturbed images ($P < .0001$). Table 1 presents the precision, recall, and fooling ratios of the model for images from the test set, overall and for each category of skin lesion. After perturbation of the images from the test set, the model was unable to classify images correctly, as all images were interpreted as Melanocytic nevi (i.e., the target category of the adversarial network).

Sensitivity and specificity of the 2 combined observers calculated on a random sample of 50 images were 58.3% (CI 95%: 45.9–70.8) and 42.5% (CI 95%: 27.2–57.8), respectively. The kappa agreement coefficient between the 2 observers was negative at −0.22 (CI 95%: −0.49–−0.04).

Figure 3 shows examples of skin lesion images from the HAM-10000 dataset before and after perturbation by the adversarial neural network.

4. Discussion

This work analyzes the impact of adversarial attacks on deep learning-based image recognition systems. Unlike previous studies, which have mostly focused on demonstrating the superior performance of deep learning algorithms compared to humans, our study explores the limits of these increasingly

Table 1

Precision, recall, and fooling ratios of the model for images from the test set, overall and for each category of skin lesion.

	Unperturbed images (n = 2003)		Perturbed images (n = 2003)		
	Precision	Recall	Precision	Recall	Fooling Ratio
Actinic keratoses	70%	65%	0%	0%	100%
Basal cell carcinoma	72%	73%	0%	0%	100%
Benign keratosis-like lesions	67%	72%	0%	0%	100%
Dermatofibroma	90%	39%	0%	0%	100%
Melanocytic nevi	91%	94%	67%	100%	NA
Melanoma	66%	58%	0%	0%	100%
Vascular skin lesions	91%	75%	0%	0%	100%
Overall	–	–	–	–	100%

NA = not applicable.

popular technologies. Our key finding is that minimal and undetectable perturbations of medical images can cause a drop in the predictive ability of image recognition algorithms. To our knowledge, only one study has focused on the demonstration of adverse attack in the field of medical algorithms.[16] This study demonstrated that adversarial attacks were capable of manipulating deep learning systems across 3 medical domains; i.e., to classify diabetic retinopathy from retinal fundoscopy, pneumothorax from chest-Xray, and melanoma from dermoscopic photographs. Although this study presented very conclusive results and the discussion was of great interest, 2 remarks on this work can be made. First, the modifications of the images were qualified as human-imperceptible on the basis of the techniques implemented, but were not tested by humans. Second, the authors are an experienced team from a prestigious university, which questions the practical feasibility of such a demonstration. Thus, it appeared necessary to confirm these results with our study.

The fact that adversarial neural network targets the image to be interpreted and not the algorithm itself means that all hospitals, radiology centers, radiology equipment manufacturers, and even medical image managers are vulnerable to this sort of attack. By extension, our study suggests the vulnerability of all deep learning applications in the field of healthcare, including for example algorithms for the clinical management of patients with sepsis or risk stratification for mortality of patients with acute myocardial infarction.[23,24]

Two opposite risks exist that raise fears of an attack like the one described here. On the one hand, there is a risk that medical databases or devices may be hacked to cause under diagnosis of a specific condition (such as skin cancer), which would result in a lack or a delay in therapeutic management. Attacks of this sort are easy to imagine—for example, for terrorist purposes or for the purpose of harming a particular person. Malicious attacks in the field of healthcare are easier to make than nuclear or bacteriological attacks, and their consequences are more difficult to predict. Moreover, while it may seem pointless to hack medical databases or devices, such attacks have in fact already occurred, for example to damage pacemakers or insulin

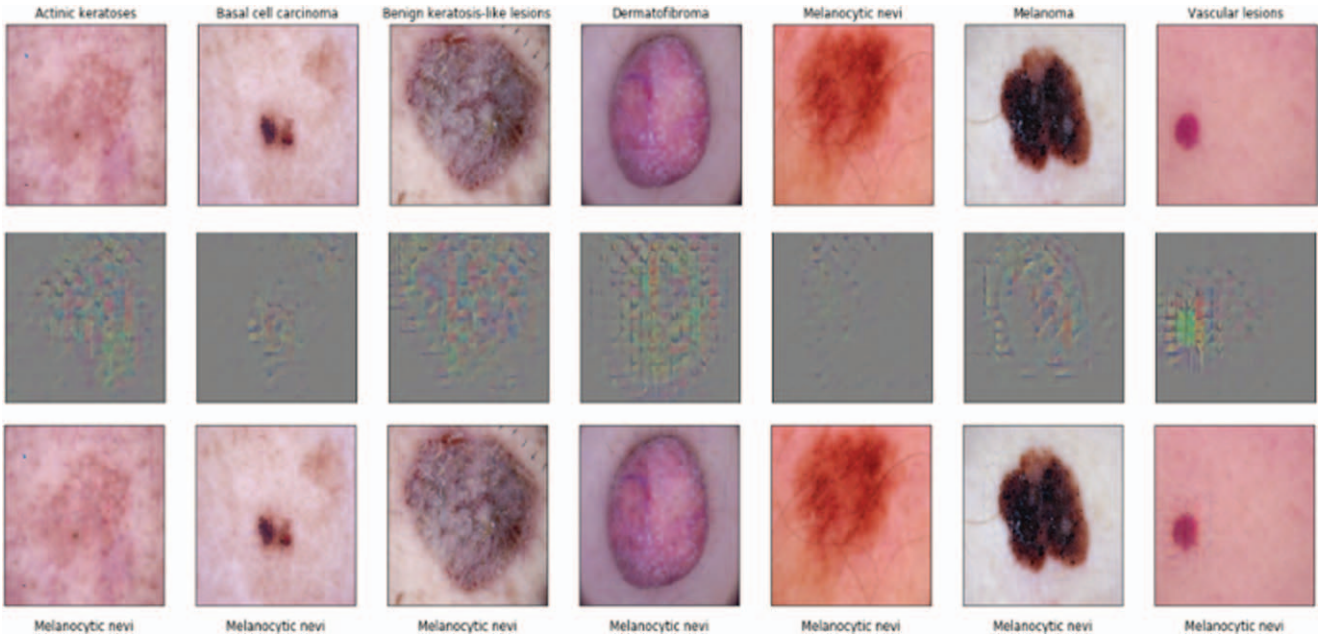

Figure 3. Examples of skin lesion images from the HAM-10000 dataset before (upper line) and after (lower line) perturbation by the adversarial neural network (resolution of 256x256 pixels).

pump systems.[25–29] On the other hand, there is a risk that medical databases or devices may be hacked to cause over diagnosis of a specific condition, which would boost the demand for care. Thus, the healthcare business sector might be tempted to tweak and falsify unverifiable data in order to generate more profit. While this may seem a far-fetched scenario, it should be noted that private actors have used industrial fraud techniques in the past—for instance, car manufacturers who wished to manipulate particle emission tests.[30,31]

As this study has shown, adversarial attacks are easy to make and do not require very strong technical expertise. The only hardware needed is a pre-trained neural network software that can easily be downloaded from the internet. Indeed, our study required less than 5 hours of human labor. Another very interesting publication has shown an alteration in the classification performances of an open access algorithm. The skin lesion image parameters were simply changed (modification of the zoom, of the adjustments of contrast/brightness settings, or rotation), and their classification thus completely distorted. The authors of this publication proposed as a possible solution the standardization of the images collected in the databases.[32]

Our study sheds light on the impact of adversarial neural network on open source or corrupted databases. Given the very real possibility of malicious attacks and the fact that databases are easily corrupted, we recommend the systematic validation of all prediction or task automation work on a different database, and not on a split of the database being used.

Solutions can be put forward to counter these malicious attacks. The first is, of course, to strengthen data security. While blockchain technology may be useful in this regard, it is associated with practical application issues.[33] The second is to develop techniques for the detection of image perturbations, as has been done in the field of facial recognition.[34] It should be recalled, however, that such techniques generally lag behind those of hackers and are therefore mostly temporary.

Our study has limitations that must be acknowledged. First, we did not try to demonstrate how images can be hacked and corrupted. This did not seem ethical to us. In any case, several malicious attacks have been made in the real world.[35,36] Second, we generated an algorithm with moderate predictive abilities (it has been found to detect melanoma with a precision of 66%). However, our aim was not to propose a more efficient algorithm —which we could easily have done by adding demographic data from the HAM-10000 dataset to our classification model. Rather, it was to highlight the drop in the predictive ability of image recognition algorithms that follows from undetectable image perturbations. In fact, far more efficient algorithms are available for use—including the algorithm proposed by Esteva et al, which is capable of classifying skin cancer with a level of competence comparable to dermatologists.[4] It is likely that these more efficient algorithms would likewise experience a drop in predictive ability should they be targeted by an adversarial attack.

5. Conclusion

In conclusion, it seems essential to improve our understanding of image recognition algorithms and to upgrade their security before putting them to practical and daily use. In addition to strengthening data protection, one could ensure that medical image recognition is subject to human control (for instance, dermatologists could be required to interpret one out of 5 skin lesion images). A legislative framework could be put in place to regulate the use of big data and artificial intelligence (AI) technology. Clinicians who use and develop AI programs could be trained in the field of cyber security. Research centers using AI for clinical purposes could be required to reassess and update their security systems on a regular basis. In other words, deep learning algorithms should be used to help humans, and not to replace them—at least for the time being.

Acknowledgments

We would like to thank Arianne Dorval for English language editing.

Author contributions

Jerome Allyn, Nicolas Allou, and Cyril Ferdynus were involved in the conception, data collection, data analysis and writing of the manuscript. Charles Vidal and Amelie Renou were involved in the conception and writing of the main manuscript text.

The datasets generated during and/or analyzed during the current study are available in the HAM-10000 repository: https://www.nature.com/articles/sdata2018161

Conceptualization: Jerome Allyn, Nicolas Allou, Amélie Renou, Cyril Ferdynus.

Formal analysis: Amélie Renou, Cyril Ferdynus.

Investigation: Amélie Renou.

Methodology: Jerome Allyn, Charles Vidal, Amélie Renou, Cyril Ferdynus.

Project administration: Charles Vidal, Cyril Ferdynus.

Supervision: Jerome Allyn, Nicolas Allou, Cyril Ferdynus.

Validation: Jerome Allyn, Charles Vidal.

Writing – original draft: Jerome Allyn, Nicolas Allou, Charles Vidal, Amélie Renou, Cyril Ferdynus.

Writing – review & editing: Jerome Allyn, Cyril Ferdynus.

References

[1] Carin L, Pencina MJ. On Deep Learning for Medical Image Analysis. JAMA 2018;320:1192–3.
[2] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.
[3] Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318:2211–23.
[4] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.
[5] Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15:e1002686.
[6] Lu D, Popuri K, Ding GW, et al. Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural MR and FDG-PET images. Sci Rep 2018;8:5697.
[7] Gandhi S, Mosleh W, Shen J, et al. Automation, machine learning, and artificial intelligence in echocardiography: a brave new world. Echocardiography 2018;35:1402–18.
[8] Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886.
[9] Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20–3.
[10] Finlayson SG, Bowers JD, Ito J, et al. Adversarial attacks on medical machine learning. Science 2019;363:1287–9.
[11] Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. arXiv.org 2017; https://arxiv.org/abs/1710.08864. [accessed June 8, 2020]
[12] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv.org 2013; https://arxiv.org/abs/1312.6199v4. [accessed June 8, 2020]

[13] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The Limitations of Deep Learning in Adversarial Settings. arXiv.org 2015; https://arxiv.org/abs/1511.07528v1. [accessed June 8, 2020]
[14] Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. arXiv.org 2015; https://arxiv.org/abs/1511.04599v3. [accessed January 8, 2020]
[15] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks. ArXiv.org 2016; http://arxiv.org/abs/1608.04644. [accessed June 8, 2020]
[16] Finlayson SG, Chung HW, Kohane IS, Beam AL. Adversarial Attacks Against Medical Deep Learning Systems. arXiv.org 2019; https://arxiv.org/abs/1804.05296. [accessed June 8, 2020]
[17] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5:1–9.
[18] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. arXiv.org 2016. https://arxiv.org/abs/1608.06993v5. [accessed June 8, 2020]
[19] Yu X, Zeng N, Liu S, et al. Utilization of DenseNet201 for diagnosis of breast abnormality. Mach Vis Appl 2019;30:1135–44.
[20] ImageNet. Available at http://www.image-net.org/. [accessed June 8, 2020]
[21] Poursaeed O, Katsman I, Gao B, Belongie S. Generative Adversarial Perturbations. ArXiv.org 2017; http://arxiv.org/abs/1712.02328. [accessed June 8, 2020]
[22] Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv.org 2016; https://arxiv.org/abs/1603.04467v2. [accessed June 8, 2020]
[23] Kwon JM, Jeon KH, Kim HM, et al. Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction. PloS One 2019;14:e0224502.
[24] Komorowski M, Celi LA, Badawi O, et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018;24:1716–20.
[25] Baranchuk A, Refaat MM, Patton KK, et al. Cybersecurity for cardiac implantable electronic devices: what should you know? J Am Coll Cardiol 2018;71:1284–8.
[26] Frenger P. Hacking medical devices a review. Biomed Sci Instrum 2013;49:40–7.
[27] VentureBeat. Insulin pump hacker says vendor Medtronic is ignoring security risk. Available at https://venturebeat.com/2011/08/25/insulin-pump-hacker-says-vendor-medtronic-is-ignoring-security-risk/. [accessed June 8, 2020]
[28] VentureBeat. Excuse me while I turn off your insulin pump. Available at https://venturebeat.com/2011/08/04/excuse-me-while-i-turn-off-your-insulin-pump/. [accessed June 8, 2020]
[29] Pycroft L, Aziz TZ. Security of implantable medical devices with wireless connections: the dangers of cyber-attacks. Expert Rev Med Devices 2018;15:403–6.
[30] Gates G, Ewing J, Russell K, et al. The New York Times. How Volkswagen's 'Defeat Devices' Worked. Available at https://www.nytimes.com/interactive/2015/business/international/vw-diesel-emissions-scandal-explained.html. [accessed June 16, 2020]
[31] Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018;378:981–3.
[32] Navarrete-Dechent C, Dusza SW, Liopyris K, et al. Automated dermatological diagnosis: hype or reality? J Invest Dermatol 2018;138:2277–9.
[33] Leeming G, Ainsworth J, Clifton DA. Blockchain in health care: hype, trust, and digital health. Lancet Lond Engl 2019;393:2476–7.
[34] Massoli FV, Carrara F, Amato G, Falchi F. Detection of Face Recognition Adversarial Attacks. arXiv.org 2019; https://arxiv.org/abs/1912.02918. [accessed June 8, 2020]
[35] Choi SJ, Johnson ME, Lehmann CU. Data breach remediation efforts and their implications for hospital quality. Health Serv Res 2019;54:971–80.
[36] McAfee Labs Threats Report. Available at https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-aug-2019.pdf. [accessed June 8, 2020]