Adversarial radiomics: the rising of potential risks in medical imaging from adversarial learning

Andrea Barucci¹ · Emanuele Neri²

Published online: 26 May 2020

© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Introduction

Radiomics is defined as a “high-throughput extraction of quantitative features that result in the conversion of images into mineable data and the subsequent analysis of these data for decision support” [1]. It is becoming a huge field of research of medical imaging and, de facto, is promising as a cornerstone in precision medicine along with the other omics science, having shown great potential in many different clinical applications [2–4].

One important step in the radiomics development is that it is progressively moving towards deep radiomics, namely the use of deep learning to automatically extract features from images, classify disease, and predict outcomes. This deep approach is probably overcoming the traditional radiomic approach using hand-crafted features (i.e., traditional radiomics) [5–7].

Despite the positive aspects and promises of a significant impact of radiomics in clinical practice, we are also aware of some risks to which this technology may be subject. For example, the lack of reproducibility in radiomic studies is a well-known problem, basically depending on the complex mixing of the many different steps in the data acquisition, processing, and analysis, translating in a sort of pipeline fingerprint which can affect the results of the analysis, but the transfer of knowledge from other fields lets us perceive that there is a more devious threat [8, 9]. How will we change our trustworthiness

in radiomics if despite taking care of all sources of data confusion, just changing a few pixels on an image (not randomly), what looks the same to the human eye causes the state of the art classifiers to fail miserably?

What about if perturbing some ad hoc pixel values in an image, we will be able to manipulate a diagnosis, resulting in a user-predefined category?

Adversarial machine learning

We have been recently advised of such potential situation from the machine learning community, where the so-called *adversarial machine learning* field is developing from years. Adversarial machine learning is a technique employed in the field of machine learning which attempts to fool models through malicious input [10]. Many studies have shown how a small (and eventually carefully designed) perturbation of the data is able to totally deceive models, independently from the class of the algorithm [11, 12].

Today, we are aware that every machine learning domain needs to face this treat, healthcare in primis, where issues related to personal, ethical, financial, and legal consequences have a huge impact on society [13–16].

In a historical period where artificial intelligence technologies are permeating many aspects of life, where concepts as trustworthy and accountability are becoming central, applications in medical imaging deserve deeper attention in order to fulfill those requirements [17, 18].

Recently, some studies on adversarial examples in clinical radiology have shown as, e.g., a benign correctly classified mole can be changed in malignant (with 100% confidence) just by adding a specifically optimized adversarial noise [13], and any quantitative imaging modality can be corrupted with specifically developed attack strategies [19–22], as in the study of Mirsky et al. [23], where authors implemented an attack using a generative adversarial network (GAN),

This article is part of the Topical Collection on Advanced Image Analysis

✉ Andrea Barucci
a.barucci@ifac.cnr.it

¹ CNR-IFAC Institute of Applied Physics “N. Carrara”, 50019 Sesto Fiorentino, Italy

² Diagnostic and Interventional Radiology, Department of Translational Research, University of Pisa, Via Roma, 67, 56126 Pisa, Italy

showing how to automate the framework and producing false lesions in computed tomographies. The authors trained the GAN with 888 annotated CT scans and performed an authorized simulation by entering a hospital's radiology department; they introduced a single-board computer, with wireless local area network and Bluetooth connectivity, and used it as a Trojan horse to access the PACs network; they were able to intercept CT scans and introduce the desired modifications.

Adversarial radiomics

Radiomics, moving medical imaging towards a data-driven approach, is then the perfect field for adversarial examples, whatever the approach chosen (traditional or deep radiomics). Although traditional radiomics may seem less subject to this kind of adversarial examples, it is not difficult to think of small modifications in the image able to suitably change the values of some features, fooling the subsequent data analysis [8].

In analogy to adversarial learning, we introduce the term *adversarial radiomics*, referring to “adversarial examples in radiomics”.

It is very important to stress the difference between adversarial radiomics and intrinsic problems of radiomics due to reproducibility [24]. While of course these phenomena can share some common sources and features, there are some basic properties that distinguish them.

While the principal source of lack of reproducibility in radiomics can be reconducted to different clinical image acquisition settings, pre- and post-processing images transformation, sparsity of data, and analysis algorithms, some of these concerns can be thought to be fixed, e.g., improving data analysis techniques or/and harmonizing the datasets in terms of protocols and processing. Instead, adversarial radiomics has its roots in the problem of adversarial examples in machine learning, that is more related to the intrinsic nature of algorithms [14]. Moreover, while standard errors in radiomics can be thought of as due to chance, adversarial learning can be guided to give precise wrong results. The idea leading adversarial examples is to find the smallest perturbation (in extremis just one pixel) able to fool the model classification, giving a wrong unpredictable or predictable result. The latter is the result of what is called a targeted attack, that is a deliberate manipulation of clinical images causing a misdiagnosis motivated by, e.g., insurance fraud, cyber terrorism, sabotage research, or in extremis stop a political candidate or even commit murder.

Adversarial examples are in general hard to defend against, machine learning models being trained only on a very small amount of all the many possible inputs they might encounter. Moreover it is difficult to construct a theoretical model of the adversarial example crafting process.

Awareness of adversarial examples gave rise to defense strategies [21], looking for algorithms resilient to adversarial attacks, able to make decisions for consistently explainable and appropriate reasons [13, 25]. Some strategies have been proposed such as using clever data processing to mitigate potential tampering or exposing algorithms to adversarial examples in the training [11, 13]. Woods et al. [25] suggest that such attacks may be leveraged to produce cogent explanations of robust networks. However, designing a defense that can protect against a powerful, adaptive adversarial attacker remains an important research area.

Conclusion

In a modern view of medicine, where big data sharing and analysis with machine learning are becoming a pillar of research and clinical practice, adversarial radiomics is then a subject of strong interest.

How should we defend against adversarial examples? First of all, becoming aware of their existence. In our vision, a possible solution could be the integration of all the quantitative and qualitative data in data analysis, achieving the complete picture of the patient and at the same time inserting a human expert in the loop giving feedback on explainable decisions. As suggested by Zhou and Firestone, human intuition can be a reliable guide to adversarial machine learning, making such integrated systems difficult to fool [26].

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278:563–77. [Internet]. <https://doi.org/10.1148/radiol.2015151169>.
2. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine [Internet]. *Nat Rev Clin Oncol*. 2017; 749–62. <https://doi.org/10.1038/nrclinonc.2017.141>.
3. Neri E, de Souza N, Brady A, Bayarri AA, Becker CD, Coppola F, et al. What the radiologist should know about artificial intelligence – an ESR white paper. *Insights Imaging*. 2019. [Internet]. <https://doi.org/10.1186/s13244-019-0738-2>.
4. Neri E, Del Re M, Paiar F, Erba P, Cocuzza P, Regge D, et al. Radiomics and liquid biopsy in oncology: the holons of systems medicine [Internet]. *Insights Imaging*. 2018;9:15–24. Springer Verlag. <https://doi.org/10.1007/s13244-018-0657-7>.
5. Dong T, Yang C, Cui B, Zhang T, Sun X, Song K, et al. Development and validation of a deep learning radiomics model predicting lymph node status in operable cervical cancer. *Front Oncol*. 2020;10:464. [Internet]. <https://doi.org/10.3389/fonc.2020.00464>.

6. Zhou H, Jin Y, Dai L, Zhang M, Qiu Y, Wang K, et al. Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images. *Eur J Radiol.* 2020;127:108992. [Internet]. <https://doi.org/10.1016/j.ejrad.2020.108992>.
7. Hosny A, Aerts HJ, Mak RH. Handcrafted versus deep learning radiomics for prediction of cancer therapy response [Internet]. *Lancet Digital Health.* 2019:e106–7. [https://doi.org/10.1016/s2589-7500\(19\)30062-7](https://doi.org/10.1016/s2589-7500(19)30062-7).
8. Zhovannik I, Bussink J, Traverso A, Shi Z, Kalendralis P, Wee L, et al. Learning from scanners: bias reduction and feature correction in radiomics. *Clin Transl Radiat Oncol.* 2019;19:33–8. [Internet]. <https://doi.org/10.1016/j.ctro.2019.07.003>.
9. Robins M, Solomon J, Hoye J, Abadi E, Marin D, Samei E. Systematic analysis of bias and variability of texture measurements in computed tomography. *J Med Imaging (Bellingham).* 2019;6: 033503. [Internet]. <https://doi.org/10.1117/1.JMI.6.3.033503>.
10. Kurakin A, Goodfellow I, Bengio S, Dong Y, Liao F, Liang M, et al. Adversarial attacks and defences competition [Internet]. *The NIPS '17 Competition: Building Intelligent Systems*; 2018. p. 195–231. https://doi.org/10.1007/978-3-319-94042-7_11
11. Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning [Internet]. *Pattern Recogn.* 2018:317–31. <https://doi.org/10.1016/j.patcog.2018.07.023>.
12. Aghdam HH, Heravi EJ, Puig D. Explaining adversarial examples by local properties of convolutional neural networks [Internet]. *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.* 2017. <https://doi.org/10.5220/0006123702260234>
13. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science.* 2019;363:1287–9. [Internet]. <https://doi.org/10.1126/science.aaw4399>.
14. Papangelou K, Sechidis K, Weatherall J, Brown G. Toward an understanding of adversarial examples in clinical trials [Internet]. *Machine Learning and Knowledge Discovery in Databases*; 2019. p. 35–51. https://doi.org/10.1007/978-3-030-10925-7_3
15. Smuha NA. The EU approach to ethics guidelines for trustworthy artificial intelligence [Internet]. *Comput Law Rev Int.* 2019:97–106. <https://doi.org/10.9785/cr-2019-200402>.
16. Ethics and artificial intelligence in public health social work [Internet]. *Artificial Intelligence and Social Work*; 2018. p. 231–49. <https://doi.org/10.1017/9781108669016.015>
17. Brady AP, Neri E. Artificial intelligence in radiology-ethical considerations. *Diagnostics (Basel).* 2020;10. [Internet]. <https://doi.org/10.3390/diagnostics10040231>.
18. Neri E, Coppola F, Miele V, Bibbolino C, Grassi R. Artificial intelligence: who is responsible for the diagnosis? *Radiol Med.* 2020. [Internet]. <https://doi.org/10.1007/s11547-020-01135-9>.
19. Taghanaki SA, Das A, Hamameh G. Vulnerability analysis of chest X-ray image classification against adversarial attacks [Internet]. *Understanding and interpreting machine learning in medical image computing applications*; 2018. p. 87–94. https://doi.org/10.1007/978-3-030-02628-8_10
20. Vatian A, Gusarova N, Dobrenko N, Dudorov S, Nigmatullin N, Shalyto A, et al. Impact of adversarial examples on the efficiency of interpretation and use of information from high-tech medical images [Internet]. *2019 24th Conference of Open Innovations Association (FRUCT)*; 2019. <https://doi.org/10.23919/fruct.2019.8711974>
21. Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J, et al. Understanding adversarial attacks on deep learning based medical image analysis systems [Internet]. *Pattern Recogn.* 2020:107332. <https://doi.org/10.1016/j.patcog.2020.107332>.
22. Paschali M, Conjeti S, Navarro F, Navab N. Generalizability vs. robustness: investigating medical imaging networks using adversarial examples [Internet]. *Medical image computing and computer assisted intervention – MICCAI 2018*; 2018. p. 493–501. https://doi.org/10.1007/978-3-030-00928-1_56
23. Mirsky Y, Mahler T, Shelef I, and Elovici Y. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning, SEC'19: Proceedings of the 28th USENIX Conference on Security Symposium, Santa Clara (CA), USA, 2019
24. Li Y, Han G, Wu X, Li Z, Zhao K, Zhang Z, et al. Normalization of multicenter CT radiomics by a generative adversarial network method. *Phys Med Biol.* 2020. [Internet]. <https://doi.org/10.1088/1361-6560/ab8319>.
25. Woods W, Chen J, Teuscher C. Adversarial explanations for understanding image classification decisions and improved neural network robustness [Internet]. *Nat Mach Intell.* 2019:508–16. <https://doi.org/10.1038/s42256-019-0104-6>.
26. Zhou Z, Firestone C. Humans can decipher adversarial images. *Nat Commun.* 2019;10:1334. [Internet]. <https://doi.org/10.1038/s41467-019-08931-6>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.