European Journal of Nuclear Medicine and Molecular Imaging (2021) 48:3046–3047
https://doi.org/10.1007/s00259-021-05417-w

EDITORIAL

Explaining a XX century horse behaviour

Noemi Gozzi[1] · Arturo Chiti[1,2]

Published online: 25 May 2021

In this issue of the EJNMMI, Weber and colleagues [1] reported an engaging experience using a prototype PET-assisted reporting system (PARS), employing a neural network to identify areas of suspicious FDG uptake. They used a neural network trained on images from patients affected by lung cancer and lymphoma, but the authors used the PARS to evaluate images from patients with breast cancer. The study revealed that PARS had high accuracy in foci delineation and anatomical position determination in breast cancer when evaluating PERCIST measurable lesions only. PARS performance was much lower when assessing all tumour foci, including those manually delineated by imaging experts. In the end, it seemed that the PARS neural network was able to identify lesions from breast cancer, although being trained on different tumour entities.

The message coming from this paper might have a significant impact on the future use of automatic image reading using neural networks.

Acknowledging the importance of this experiment, we would like to comment on how "knowledge" acquired by a neural network can be used extensively and mention the possible problems related to this algorithm's "behaviour".

"Clever Hans" owes its name to an experiment conducted in the 1900s. Clever Hans was an Orlov Trotter horse that was considered a scientific sensation due to its supposed ability to perform arithmetic tasks [2]. The psychologist Oskar Pfungst demonstrated that Hans did not master math; Hans was observing the reactions of the trainer and deriving the correct answer from cues in his body language. Indeed, if Hans could not see the trainer's body cue, he could not resolve the same questions.

"Clever Hans" phenomena are often applied in artificial intelligence studies to indicate predictors and models that learn spurious correlations in the training data. Similarly to Clever Hans horse, these models seem able to correctly perform a task, while they do not understand the underlying meaningful patterns in the data. It can occur when a feature in the input data correlates significantly with the outcome, like the trainer's body language correlated with the correct arithmetic result. However, there is no true causality between this feature and the correct output of the model. These models will likely fail in the real world, where these spurious correlations are not present.

A meaningful example is reported in a paper by Ribeiro et al. [3]. A state-of-the-art neural network (Google's pre-trained Inception neural network) extracts features from images that then input in a logistic regression classifier to distinguish huskies and wolves. The models were trained on biased datasets where all wolves' images had a snow background, while pictures of huskies did not. This experiment showed that the models learned to predict "Wolf" if there was snow and "Husky" otherwise, instead of learning meaningful patterns in the input data such as animal characteristics or colour and texture of the fur. The algorithm could not correctly identify the animals in a real-world dataset with huskies and wolves with a random background. Similarly, Lapuschkin et al. [4] reported that the algorithms learned to classify trains and boats using the PASCAL VOC 2012 dataset (https://deepai.org/dataset/pascal-voc) recognising rails and water, respectively. In an actual application such as object recognition and localisation for autonomous driving, this model would not have recognised a boat on a trailer on the road or a train when the rails are not visible, leading to problematic consequences.

Evidence of Clever Hans predictors also exists in clinical applications. Winkler et al. [5] investigated the association of surgical skin markers with benign nevi/malignant melanoma classification in dermoscopic images. A pre-trained convolutional neural network (Inception-v4), trained with more than 120,000 dermoscopic images, was independently

This article is part of the Topical Collection on Advanced Image Analyses (Radiomics and Artificial Intelligence)

✉ Arturo Chiti
arturo.chiti@hunimed.eu

[1] IRCCS Humanitas Research Hospital, Via Manzoni 56, Rozzano, 20089 Milan, Italy

[2] Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, Pieve Emanuele, 20090 Milan, Italy

tested on three image sets with 130 melanocytic lesions. The work included three different scenarios: (i) original marked dataset (some images with gentian violet skin markers), (ii) unmarked dataset where skin markers were digitally removed from images, and (iii) dataset with cropped images, used to remove skin markers. The CNN achieved a sensitivity of 95.7% and 100% and a specificity of 84.1% and 45.8% for unmarked and marked datasets. Marked datasets increased malignant melanoma probability score; using heatmaps created by vanilla gradient, the increase in false positive correlated with gentian violet skin markers in the images. With a predominance of skin markers in malignant melanoma in the training set, a dataset bias may have affected the learning process and induced the model's association of surgical skin markers with malignant melanoma. To further demonstrate that this association was solely due to the dermoscopic background and not to lesions' characteristics, the test on cropped images achieved the best diagnostic performance with a sensitivity of 100% and a specificity of 97.2%.

Clever Hans predictors are significant obstacles in the AI transition from scientific research to routine clinical applications. A recent paper describes the issue: "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans" [6]. Roberts et al. discussed the potential clinical use of machine learning and deep learning methods proposed in the literature for COVID-19 diagnosis and prognosis. None of the works under review was considered reliable enough to be applied in clinical routine due to methodological flaws, dataset biases, poor integration of multimodal data, and lack of reproducibility. Insufficient quality training data leading to Clever Hans predictors is a significant weakness in developing and accepting AI-powered medicine in clinical routine, which is the endpoint of scientific research in the field.

To address the possible shortcomings related to the use of biased datasets, we can think about undertaking some improvement measures, as proposed below:

1. Use data coming from studies with a robust design, avoiding biases and unbalances concerning gender, age, and other demographic characteristics and with reliable well-defined ground truth (reference standard).

2. Train algorithms employing large and variable datasets from multi-institutions collections, with a correctly defined data fusion and integration among institutions, and an external independent test set for model's validation to avoid overfitting performance. Avoid using "Frankenstein datasets" [6], made up from pieces of other datasets.

3. Researchers should use ante-hoc data investigation (e.g. exploratory data analysis with frequency tables) and post-hoc classification analysis, like explainable artificial intelligence (XAI) [7] and interpretability methods, to verify the correctness of the dataset and the classification outcomes.

The application of artificial intelligence in medical imaging is one of the fascinating changes we are leaving in these years. It is of paramount importance that we understand the limitation and pitfalls of the so-called "artificial intelligence algorithms" to take the best advantage of these tools. We should do our best to generate reliable data to train algorithms and make their conclusion explainable and reliable. In this effort, the cooperation between scientists from different disciplines is an essential requirement for success.

Declarations

Ethics approval Institutional Review Board approval was not required because the paper is an Editorial.

Informed consent Not applicable.

Conflict of interest The authors declare no competing interests.

References

1. Weber M, Kersting D, Umutlu L, et al. Just another "Clever Hans"? Neural networks and FDG PET-CT to predict the outcome of patients with breast cancer. Eur J Nucl Med Mol Imaging. 2021. https://doi.org/10.1007/s00259-021-05270-x.
2. Pfungst O (1911) Clever Hans (The horse of Mr. von Osten): A contribution to experimental animal and human psychology (Trans. CL Rahn). New York: Henry Holt. (Originally published in German, 1907)
3. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier," CoRR, vol. abs/1602.04938, 2016. http://arxiv.org/abs/1602.04938
4. Lapuschkin S, Binder A, Montavon G, Müller K, Samek W. Analyzing classifiers: fisher vectors and deep neural networks. 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2912–2920. https://doi.org/10.1109/CVPR.2016.318.
5. Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol. 2020;10:1135–41. https://doi.org/10.1001/jamadermatol.2019.1735.
6. Roberts MT, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3:199–217. https://doi.org/10.1038/s42256-021-00307-0.
7. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fus. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.