Interpretable Artificial Intelligence: Why and When

Adarsh Ghosh[1]
Devasenathipathy Kandasamy

OBJECTIVE. The purpose of this article is to discuss the problem of interpretability of artificial intelligence (AI) and highlight the need for continuing scientific discovery using AI algorithms to deal with medical big data.

CONCLUSION. A plethora of AI algorithms are currently being used in medical research, but the opacity of these algorithms makes their clinical implementation a dilemma. Clinical decision making cannot be assigned to something that we do not understand. Therefore, AI research should not be limited to reporting accuracy and sensitivity but, rather, should try to explain the underlying reasons for the predictions, in an attempt to enrich biologic understanding and knowledge.

Recent regulations introduced in the European Union require that artificial intelligence (AI) algorithms that depend on user-level predictors to make decisions provide explanations, especially when the outcomes have a significant effect on personal outcome. Utilization of AI, especially deep learning research, is increasing in radiology, pathology, and medicine in general. However, because such algorithms affect patient outcomes, the black box–like structure of deep learning algorithms remains a pet peeve [1]. We do not know exactly how the algorithms work, and therefore we cannot anticipate when the algorithms will fail. The purported advantages of AI include a potential reduction in health care costs and human resources requirements as well as freedom from human bias. AI is thus predicted to serve as a clinician assistant that can handle various tasks, including determining the response to chemotherapy, interpreting images, and choosing therapeutic regimens, among other tasks. Given the promise of AI in the routine clinical management of patients, it is imperative that we understand the workings of the algorithms being trained rather than use them as a black box and not understand how a decision is made. “Interpretability is the degree to which a human can understand the cause of a decision” [2] made by an algorithm. Interpretability of any new method is paramount in the medical field. Similar to the various checks and balances applied to the clinical use of any pharmaceutical agent, a discussion of interpretability should be paramount in AI research published in medical journals.

One might argue against interpretability. A common refrain is that if AI can do better than humans, then why not use AI, because even humans can make mistakes. However, when a clinician or a radiologist makes a mistake, fact-finding missions can often underscore the reason for the mistake. A surgeon cannot justify performing a radical pelvic lymph node dissection just because an opaque model using image-based parameters predicted deep myometrial invasion [3]. A typical comeback to this is “Why can a surgeon perform a radical pelvic lymph node dissection when a radiologist predicts deep myometrial invasion but not when a machine predicts the same finding?” The answer to this question lies in interpretability. When a radiologist predicts myometrial invasion, the prediction is based on tangible findings from imaging that even the surgeon can see at the time of surgery and that can be confirmed by the pathologist by microscopy. An opaque AI algorithm, on the other hand, does not lend itself to evaluation of how the interplay of radiomics or imaging features provided the algorithm’s prediction. Such opacity hinders clinical implementation, no matter how accurate the models are.

Another lacuna in the literature on machine learning is that it not add to scientific

Keywords: biomedical research, deep learning, machine learning

doi.org/10.2214/AJR.19.22145

Received August 13, 2019; accepted after revision October 11, 2019.

Based on a presentation at the Radiological Society of North America 2019 annual meeting, Chicago, IL.

[1] Both authors: Department of Radiodiagnosis, All India Institute of Medical Sciences, Ansari Nagar, New Delhi, 110029, India. Address correspondence to D. Kandasamy (devammc@gmail.com).

AJR 2020; 214:1137–1138

ISSN-L 0361–803X/20/2145–1137

understanding. The ultimate aim of science is to bring forth the unknown using hypothesis and rebuttals. The predictions obtained from AI algorithms cannot be the end of AI research, and it is crucial that AI modeling moves from a probabilistic approach to uncovering underlying causal relations. Most AI research uses a multitude of clinical or imaging parameters to predict a clinical outcome like recurrence or survival. Most of these studies limit themselves to comparing the accuracies of the developed models, so no new biologic insight is obtained regarding the interplay of the various parameters to provide the outcome. Although machine learning is a very convenient method of exploring big medical data, researchers and peer reviewers should not limit themselves to accuracy-driven metrics and should attempt to the explore the concrete biologic explanations underlying the opaque models being built. In the long run, this will enable medical discovery. Let us consider, for example, a study by Bae et al. [4] that involved radiomic MRI phenotyping of glioblastomas. Bae and colleagues modeled various radiomic parameters using the random survival forest algorithm to predict overall survival and progression-free survival. Although they conveyed the importance of features, their emphasis was on the predictive accuracy of the models. Radiomics forms the basis of big data in radiology, and machine learning algorithms are a very convenient method of exploring data. However, the models built by the research of Bae and colleagues are opaque, and no new information is provided regarding the interplay between these various important radiomic parameters in the models and how the most essential parameters affected survival. These interpretations are meaningful because by further probing these random survival models, will we be able to find concrete explanations for radiomic phenotyping and their underlying correlations. In the long run, this will enable their use as quantitative imaging markers overcoming the limitations of black box predictive models.

Similarly, understanding how deep learning networks view images is also crucial. Let us consider, for example, a deep learning algorithm for predicting the degree of liver fibrosis using gadoxetic acid–enhanced hepatobiliary phase images [5]. Visual interpretation of liver fibrosis on routine MR images is highly variable and inaccurate; however, radiologists have classically graded liver fibrosis as mild, moderate, or severely coarse on ultrasound. These same images are, however, interpreted more accurately by deep learning algorithms. Thus, rather than just describing the algorithms and their accuracy in predicting fibrosis, further scientific discovery should be directed at identifying which aspects of the images were used by the algorithm for evaluating fibrosis. This approach will not only satiate our hunger for knowledge but may also provide radiologists with visual cues for identifying fibrosis and predicting it with a better degree of accuracy than is achieved using current visual standards. The local interpretable model-agnostic explanations (LIME) method is one such method that can help us evaluate the black box and understand how images were interpreted. With use of this method, the image in question is perturbed in different ways and then fed into the black box of the algorithm, and the prediction is subsequently analyzed. Analysis of the image perturbations will allow us to interpret which aspect of the image actually led to a benign or malignant classification. It would be an interesting exercise to evaluate which imaging features were picked up by the algorithms to classify solitary pulmonary nodules accurately. Such an evaluation would enrich radiologic interpretation and might identify a new imaging marker of malignancy. Use of such a method may also allow review of hundreds of thousands of images to assess the imaging features of malignancy, adding great strength to the descriptive methods currently used.

That an AI algorithm can go horribly wrong has been established. For example, the livestreaming of a terrorist attack in New Zealand on various social networking sites was seen as a significant failure of the AI deep learning algorithms used to filter such content. The cause cited for this failure was the absence of extensive training data of a similar nature. Given the fact that exception is almost always a rule in medicine, black box models cannot be used in the clinical sciences. For example, the decision tree used in a study by Khalaf et al. [6] presents a precise and very interpretable model for predicting the severity of postembolization syndrome after transarterial chemoembolization using a variety of clinical parameters. Therefore, although researchers may be tempted to use opaque algorithms in the pursuit of higher accuracies, a simple decision tree may provide more significant insights into the underlying biology and, along the way, may avoid endangering patients through the malfunctioning of opaque algorithms.

According to Doshi-Velez and Kim [2], interpretable algorithms should be fair, reliable, and robust and should protect data privacy. Interpretability is neither definitive nor needed for all algorithms. For example, an AI algorithm for image segmentation or motion correction would not require transparency. Interpretability will not be the most pressing concern because the results would have a predominant research or data mining application. On the other hand, if these algorithms are used to make treatment decisions independent of the clinician, interpretability becomes imperative.

Conclusion: The Way Forward

Reporting the accuracy of AI algorithms ideally should not be the endpoint of AI research. Instead, AI should be used for better evaluation and visualization of big data currently making inroads into medical research. Scientific discovery should remain the main driving force behind research published in medical and radiology journals, and AI research should not be limited to reporting accuracy and sensitivity compared with those of the radiologist, pathologist, or clinician. More importantly, reports of AI research should try to explain the underlying reasons for the predictions, in an attempt to enrich biologic understanding and knowledge.

References

1. Ford RA, Price W, Nicholson I. Privacy and accountability in black-box medicine. *Michigan Telecommunications and Technology Law Review* 2016; 23:12–19
2. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv website. arxiv.org/abs/1702.08608. Revised March 2, 2017. Accessed August 12, 2019
3. Ueno Y, Forghani B, Forghani R, et al. Endometrial carcinoma: MR imaging-based texture model for preoperative risk stratification—a preliminary analysis. *Radiology* 2017; 284:748–757
4. Bae S, Choi YS, Ahn SS, et al. Radiomic MRI phenotyping of glioblastoma: improving survival prediction. *Radiology* 2018; 289:797–806
5. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid-enhanced hepatobiliary phase MR images. *Radiology* 2018; 287:146–155
6. Khalaf MH, Sundaram V, AbdelRazek Mohammed MA, et al. A predictive model for postembolization syndrome after transarterial hepatic chemoembolization of hepatocellular carcinoma. *Radiology* 2018; 290:254–261