OXFORD

JNCI J Natl Cancer Inst (2019) 111(9): djy226

doi: 10.1093/jnci/djy226
First published online January 10, 2019
Editorial

EDITORIAL

Shining Light Into the Black Box of Machine Learning

William Hsu, Joann G. Elmore

See the Notes section for the authors' affiliations.
Correspondence to: Joann G. Elmore, MD, MPH, David Geffen School of Medicine, UCLA, Los Angeles, CA (e-mail: jelmore@mednet.ucla.edu).

"By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it." —Eliezer Yudkowsky (1)

Advances in machine learning (ML) have brought artificial intelligence (AI) and its biomedicine applications into the spotlight as part of a larger and (we hope) smarter approach to many health-care challenges. Driving these advancements has been the availability of large and diverse datasets such as electronic and personal health records, disease registries, imaging and genomic repositories, and wearable sensors. Studies are showing that AI/ML algorithms can outperform humans at various diagnostic tasks from detecting polyps to diagnosing cancer (2). However, as we transition from models that are proofs-of-concept to decision support tools that affect patient care, both AI/ML developers and the end users who interact with them need to fully appreciate how and what these models are "learning."

In this issue, Hu et al. (3) describe a supervised, deep learning-based approach to predicting cervical precancers and cancers. Briefly, deep learning attempts to discover the hidden structure of complex, high dimensional data inputs using a hierarchical network consisting of multiple layers (where each layer is a collection of nodes in the network that operate together). Higher layers in the hierarchy are defined using features from lower layers. This approach has gained enormous traction in a variety of applications in health care (4). Using a retrospective dataset of 9406 women who underwent cervical cancer screening using photographic images of the cervix (cervicography), the authors trained a variant of a deep learning method called the Faster R-CNN (Region-based Convolutional Neural Network). The Faster R-CNN approach first transforms an image into a convolutional feature map using a CNN. The feature map is processed by a region proposal network, which generates regions containing objects of interest. These regions are reshaped into a fixed size and fed into a Fast R-CNN detector, which predicts the label of the inputted region. The authors' algorithm achieves better accuracy in predicting precancer/cancer compared with the original physician readers who interpreted the cervicography ($P < .001$) and better accuracy when compared with other screening tests such as conventional cytology ($P < .001$). The study highlights the power of AI/ML techniques when trained on a large, diverse dataset. However, the nuanced methodological issues that affect model generalizability (eg, the ability to reliably use the model on different cohorts) are not always reflected in summary statistics and *P* values.

A wide variety of AI/ML algorithms exist, including supervised, unsupervised, and reinforcement learning approaches. Hu et al. used supervised learning, which uses labeled information (eg, the known outcome of a patient, a region of interest drawn by a physician annotator on an image) as "truth" during model training to elucidate the hidden relationships among the input data and output labels. Algorithms in this class include logistic regression, support vector machines, and random forests. Unsupervised learning requires no predefined labels but rather uses a predetermined measure of similarity or distance to discover inherent structure within data. Algorithms that perform dimensionality reduction (eg, modeling the input dataset using fewer features) and clustering (eg, finding groups of similar features) are examples. Reinforcement learning attempts to recommend an action (eg, perform a diagnostic test) that achieves a desired outcome (eg, arrive at a definitive diagnosis for a given patient) by observing the outcome of the action (eg, whether the diagnostic test reveals the correct, definitive diagnosis). The outcome is a quantified value called a reward, which is positive in situations with good outcomes and negative in situations with poor outcomes. By observing cumulative rewards for a given action over time, reinforcement learning algorithms use this feedback to determine the optimal sequence of actions that result in the best long-term strategy. Reinforcement learning is widely being used to enable autonomous driving of cars and is increasingly being explored to suggest interventions such as when and how to prompt an individual to take a medication or to exercise (5).

A longstanding question for AI/ML developers and end users alike is: What AI/ML approach is best suited for my prediction

Received: November 28, 2018; **Accepted:** December 3, 2018

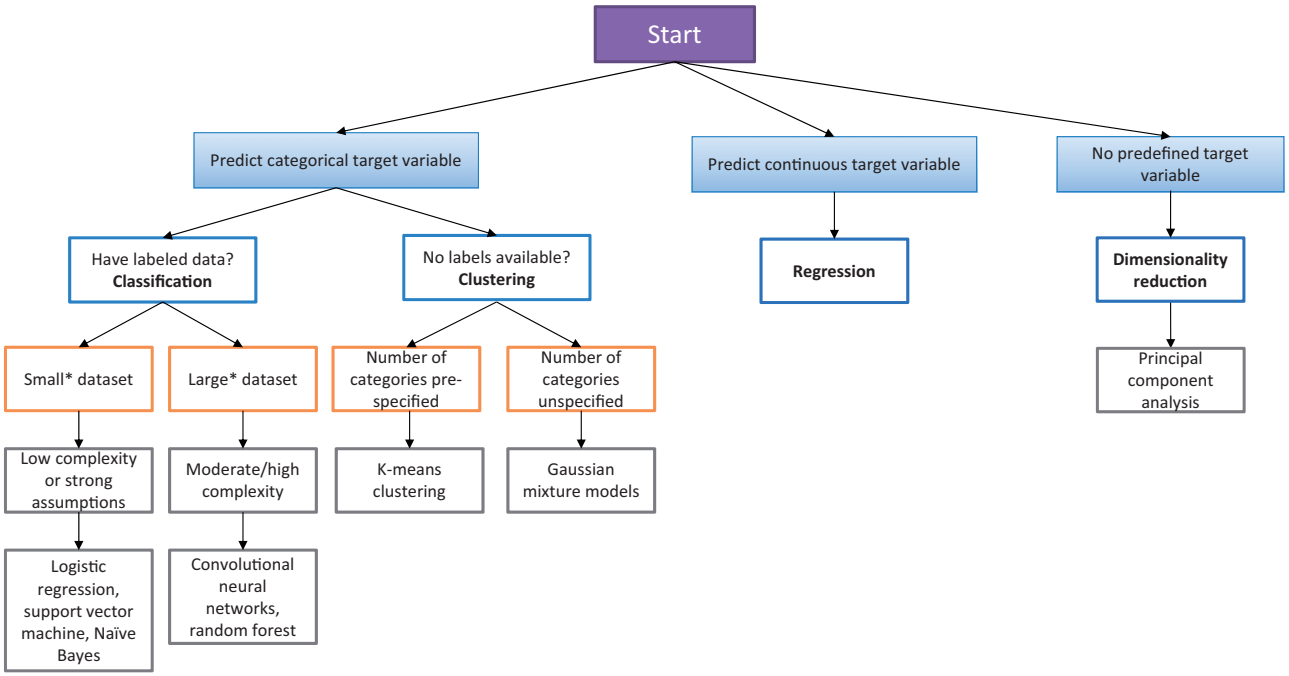

Figure 1. A simplified decision tree illustrating one approach for selecting appropriate machine learning algorithms. *Dataset sizes are not defined as fixed thresholds because the number of data points is often relative to the target variable and quality of data being used to train the model. Adapted from https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html.

task? Selecting the appropriate model is frequently nuanced but may be guided by a series of questions. What is being predicted? Is the target or outcome variable that is being predicted categorical (eg, binary or multiple distinct labels such as tumor stage) or continuous (eg, a numerical value such as the number of days to recurrence)? What is the dimensionality of the dataset (eg, the number of input features) being used to train the model? Are there a handful of data points, or do they number in the hundreds or thousands? What is the sample size and quality of the dataset? How many cases have missing values? Does the sample size support training a model with the number of predictors being included? Figure 1 illustrates a basic decision aid that takes these criteria into consideration in selecting the appropriate AI/ML modeling approach.

Developing generalizable AI/ML models is dependent on the quality and representativeness of the data used to train the model. Hu et al. (3) had access to 59 634 cervical images from over 9000 patients as part of the Guanacaste natural history study, which was a large public health study undertaken in Costa Rica. Participants in this study underwent cytology and cervicography followed by colposcopy if the patient had an abnormal screening result. However, the dataset was highly imbalanced: the number of cases (eg, patients ultimately diagnosed with cervical precancer/cancer) was far less than those who were normal (279/9406, 3.0%). Although this reflects the natural proportion of cases as often seen in cancer screening programs, training these models requires a sufficient number and diversity of examples to distinguish between precancer/cancer and noncancer features. To compensate, Hu et al. (3) used transfer learning where instead of learning the model from scratch using their data, they used the first four layers and weights from an existing model trained on 1.28 million color photographs of over 1000 types of objects (eg, animals, instruments, cars) and their labels called ImageNet (6). They then updated the values of the remaining layers using data from the Guanacaste study. Data augmentation was also performed to generate additional event cases by randomly flipping, rotating, and shearing the images as well as varying the brightness of the images as a means of simulating variability within the data. Although these strategies have been shown to improve model generalizability, the simulated cases are highly correlated with one another and may not reflect the range of possible precancer/cancer presentations.

The ability to understand models and their predictions, particularly as they model increasing numbers of data inputs, is critical. Techniques for characterizing data and model complexity such as dimensionality reduction can be applied. For models that are classifying on pixel data, visualizations such as a heatmap-like class activation map are used to visualize what regions of an image the model finds to be most informative in predicting a given class (eg, precancer/cancer vs no cancer). Hu et al (3) use class activation maps to understand model attention in correct and incorrect cases, providing a general sense of regions of the image that contributed to the prediction.

Before AI/ML models can make the transition from proof-of-concept to clinically useful algorithms, we must learn how to make the models more generalizable and understandable. As Yudkowsky's quote captures eloquently, users of AI/ML models must not see them as a "black box" but rather seek greater transparency from model developers about the inherent quality of the dataset used to train the model, the rationale behind choosing the model's representation, and the explanation associated with a model's prediction.

Notes

Affiliations of authors: Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA (WH);

Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA (JGE).

The authors have no direct conflicts of interest to disclose. Dr Elmore serves as Editor-in-Chief for Adult Primary Care at UpToDate. Dr Hsu holds a research grant from Siemens Healthineers.

References

1. Yudkowsky E. Introduction. In: Nick Bostrom and Milan M. Ćirković, eds. *Artificial Intelligence as a Positive and Negative Factor in Global Risk. Global Catastrophic Risks*. New York: Oxford University Press; 2008:308–345.
2. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500–510.
3. Hu L, Bell D, Antani S, et al. An observational study of deep-learning and automated evaluation of cervical images for cancer screening. *J Natl Cancer Inst*. 2019;111(9):923–932.
4. Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA*. 2018;320(11):1101–1102.
5. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med*. 2018;52(6):446–462.
6. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009); June 20–25, 2009; Miami, FL.