ANALYSIS

Problems in the deployment of machine-learned models in health care

Joseph Paul Cohen PhD, Tianshi Cao MSc, Joseph D. Viviano MSc, Chin-Wei Huang MSc, Michael Fralick MD PhD, Marzyeh Ghassemi PhD, Muhammad Mamdani MPH PharmD, Russell Greiner PhD, Yoshua Bengio PhD

■ Cite as: *CMAJ* 2021 September 7;193:E1390. doi: 10.1503/cmaj.202066; early-released August 30, 2021

CMAJ Podcasts: author interview at www.cmaj.ca/lookup/doi/10.1503/cmaj.202066/tab-related-content

See related articles at www.cmaj.ca/lookup/doi/10.1503/cmaj.202434 and www.cmaj.ca/lookup/doi/10.1503/cmaj.210036

In a companion article, Verma and colleagues discuss how machine-learned solutions can be developed and implemented to support medical decision-making.[1] Both decision-support systems and clinical prediction tools developed using machine learning (including the special case of deep learning) are similar to clinical support tools developed using classical statistical models and, as such, have similar limitations.[2,3] A model that makes incorrect predictions can lead its users to make errors they otherwise would not have made when caring for patients, and therefore it is important to understand how these models can fail.[4] We discuss these limitations — focusing on 2 issues in particular: out-of-distribution (or out-of-sample) generalization and incorrect feature attribution — to underscore the need to consider potential caveats when using machine-learned solutions.

Key points

- Decision-support systems or clinical prediction tools based on machine learning (including the special case of deep learning) are similar to clinical support tools developed using classical statistical models and, as such, have similar limitations.
- If a machine-learned model is trained using data that do not match the data it will encounter when deployed, its performance may be lower than expected.
- When training, machine learning algorithms take the "path of least resistance," leading them to learn features from the data that are spuriously correlated with target outputs instead of the correct features; this can impair the effective generalization of the resulting learned model.
- Avoiding errors related to these problems involves careful evaluation of machine-learned models using new data from the performance distribution, including data samples that are expected to "trick" the model, such as those with different population demographics, difficult conditions or bad-quality inputs.

What are the features of machine-learned models?

Herein we use the term "machine-learned model" to refer to a model that has been created by running a supervised machine learning algorithm on a labelled data set. Machine-learned models are trained on specific data sets, known as their training distribution. Training data are typically drawn from specific ranges of demographics, country, hospital, device, protocol and so on. Machine-learned models are not dynamic unless they are explicitly designed to be, meaning that they do not change as they are used. Typically, a machine-learned model is deterministic, having learned a fixed set of weights (i.e., coefficients or parameters) that do not change as the model is run; that is, for any specific input, it will return the same prediction every time. Although "adaptive systems" have been developed that can "learn" while being deployed by incorporating new data, such systems may give a different prediction for the same input and their safety and oversight is still unclear.[5]

We refer to the data that a machine-learned model will encounter when it is deployed for use as the model's performance distribution. If a machine-learned model's training distribution does not match its performance distribution, then the performance of the model may be lower than expected[6,7] — a challenge that is commonly referred to as out-of-distribution generalization (discussed in detail below). Another challenge is if the training data contain features that are spuriously correlated with the outcomes the tool is being designed to predict, as this may cause a machine-learned model to make predictions from the "wrong" features (also discussed below). A model's creator should seek a training data distribution that matches the performance distribution as closely as possible, and clinicians who use the tool should be aware of the exact limitations of the model's training distribution and potential shortcomings.

What are some potential problems of machine-learned models?

Out-of-distribution generalization

Newly graduated physicians are typically most comfortable managing patients who exhibit conditions they encountered during their residency training, but they are also able to manage patients with conditions they have not previously seen because

© 2021 CMA Joule Inc. or its licensors

they can use theoretical knowledge to recognize patterns of illness. In contrast, machine-learned methods are limited by the data provided during the training and development phase. Furthermore, machine-learned models do not typically know their own limits unless components are included to help the model detect when data it encounters are out of distribution (for example, a component may be built in that prevents a human chest radiograph diagnostic system from processing a photo of a cat and diagnosing pneumonia[8] — see strategies listed below). Three categories of out-of-distribution data,[9] summarized in Figure 1, include the following:

- Data that are unrelated to the task, such as obviously wrong images from a different domain; for example, magnetic resonance images presented to a machine-learned model that was trained on radiograph images; and less obviously wrong images, such as a wrist radiograph image processed using a model trained with chest radiographs
- Incorrectly prepared data; for example, blurry chest radiograph images, those with poor contrast or incorrect view of the anatomy, images presented in an incorrect file format or improperly processed, and images arising from an incorrect data acquisition protocol
- Data not included in the training data owing to a selection bias; for example, images showing a disease not present in the training data or those arising from a population demographic not similar to that of the training data set

A machine-learned model will perform suboptimally or deliver unexpected results on out-of-distribution data.

Many strategies have been developed to detect and prevent out-of-distribution data from being processed. A typical approach is for a model to compute the degree to which a data sample matches the model's training distribution, which may be presented as a score. If the score is above a certain threshold, then the model can decide not to process a data sample. One way for the model to do this — in the case of image interpretation — is for the model to attempt to reconstruct the image and compare the reconstruction to the original by some measure of similarity, such as the absolute pixel difference.[8,10] Typically, a model will do a poor job of reconstructing an image it did not encounter in training. If the reconstructed image is scored as similar enough to be judged "correct," the model can proceed to process that image; if not, processing will not occur. However, in order to build and evaluate such out-of-distribution detection systems, known out-of-distribution examples must be used; so, even strategies to prevent errors have limits.

Incorrect feature attribution

Machine-learned models typically use only the minimally complicated set of features required to reliably discriminate between the target outputs in their training data set. That is, the model takes a "path of least resistance" during its learning,[11–13] finding features that are highly predictive of the target output, which helps to make it accurate. However, a learning model may also find some distractor feature in the data that is spuriously correlated with the target output[14] and, once this happens, the model may stop looking for new true discriminative features even if they exist.[15] For example, in a model learning to read chest radiographs, distractor features may be the hospital, image acquisition parameters, radiograph view (e.g., anteroposterior v. anteroposterior supine), and artifacts such as presence of a pacemaker or endotracheal tube. If clinical protocols or image processing change over time, this can lead to patterns in the training data that can be detected by the model and serve as a distractor.[16] Or if images from multiple hospitals are grouped together and the rate of a disease varies among hospitals, a model may learn to detect the hospital using subtle visual cues and may then base its predictions on the hospital associated with the image rather than data in the image itself. This can lead

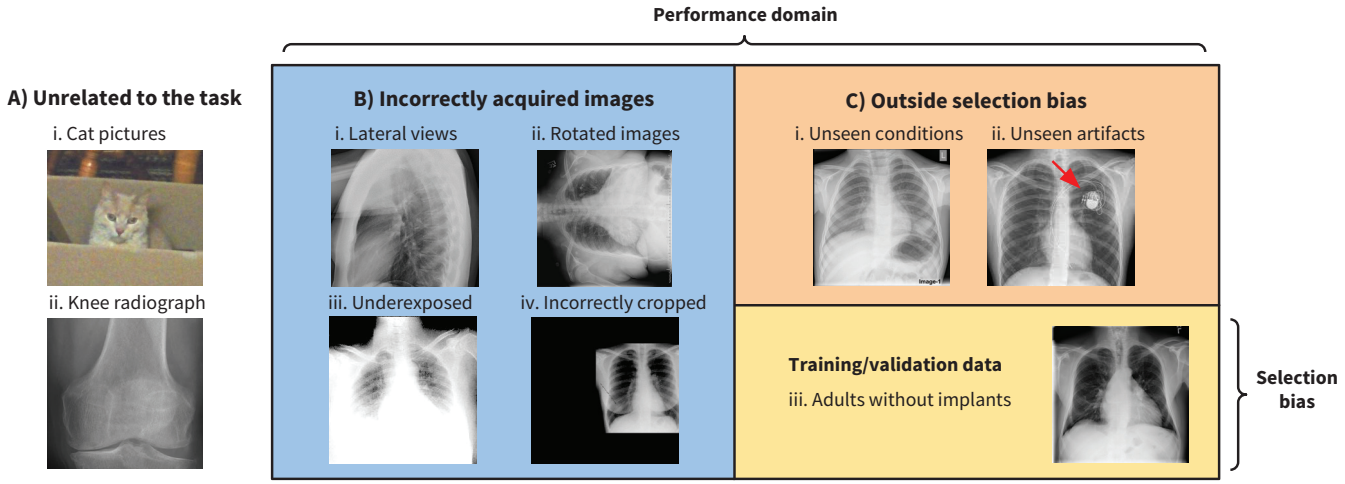

Figure 1: This figure shows 3 categories of out-of-distribution data, all in the context of training a machine-learned algorithm to read adult chest radiographs (see image C iii). A) Images that are unrelated to the task. B) Images that are incorrectly acquired. C) Images that are not encountered owing to a selection bias in the training distribution (e.g., images with lung cancer lesions and pacemakers were not included in the training set and therefore were unseen during training). C) (iii) Training data that are subject to a selection bias.

to a model appearing more accurate than it actually is if the evaluation data contain the same artifacts (e.g., the same hospital-specific distribution), but the same model could fail dramatically if the performance data do not exhibit these artifacts. Furthermore, patient demographics (e.g., age or sex) can be inferred from aspects of the training data and may be used by a learning model to predict outcome prevalence (that is, prior probability) in the training sample if better true features related to the outcome of interest are less obvious in the data.

Medical data sets are often relatively small, which may increase the likelihood of spuriously correlated features. Research into altering the ways models learn to avoid this problem is ongoing.[11,17] However, using a large, diverse data set for training a machine-learned model will help to avoid the effect of distractors. Other solutions include unsupervised learning and transfer learning,[18] processes that use data that are unlabelled or labelled for another task to train models, to avoid detection of spurious features that are specific to a particular data set. These methods typically enable the use of much more data and have a better chance of learning features that will be general enough and useful for the intended task.[18]

In cases where pathology-specific features are simply not predictive enough for some images, the learning model may be forced to guess and predict the prevalence of a disease or outcome in the training distribution. The machine-learned model will appear to work when applied to data in which the disease or outcome prevalence is the same as in the training data; it may give the "right" answer. However, when applied to a different population with a different outcome prevalence, the model will likely predict incorrectly[19,20] and lead to harm. It is therefore important that model developers and users verify that the machine-learned model appropriately detects features that are truly associated with the prediction or outcome of interest, using a feature attribution method such as the "image gradient" method[21] or creating a counterfactual input showing what would change the classifier's prediction[22] during development and when deployed.

Related to this point, another concern is that some models may simply learn to copy the actions taken by the clinicians when the data were generated. For example, if a model is trained to predict the need for blood transfusions based on historical data about transfusions, it may not have anything informative to predict from and instead will learn to replicate existing practices. A model will learn "bad habits" unless the data set used to develop it is corrected. One approach to overcome this problem would be to have expert reviewers label the data set with the true outcomes of interest (e.g., appropriate v. inappropriate blood transfusions), although this may be resource intensive and experts may not always agree on labels. It would be even better to use only labels that are objective and do not depend on human experts.

What can mitigate these problems?

Avoiding errors related to the issues discussed above involves careful evaluation of machine-learned models[23] using new data from the performance distribution, including samples that are expected to expose model failures, such as those with different population demographics, difficult conditions, poor-quality images, or errors. A potentially useful approach is to create simulated test distributions by balancing data based on attributes unrelated to the target task to observe differences in performance of a model according to factors such as demographic minority class[24] or geographic region.[25] If a model learned to focus on a spurious feature such as age, deploying it using data in which the age of the population composed of a single age, although balanced in terms of the target variable the model was trained to predict, would lead to poor performance. Results of such tests of a model's performance should be transparently presented to illustrate its limitations in use.[26] A related article discusses evaluation of machine-learned models in some depth.[27]

Conclusion

It is important to understand and tackle these problems of machine-learned models before deployment so that large investments do not end in failure, which could be costly or catastrophic. IBM's "Watson for Oncology" program[28] was suspended after an investment of $62 million, allegedly owing to problematic clinical recommendations that resulted in poor acceptance by clinicians. Google's machine-learned initiative to detect diabetic retinopathy[29] struggled when it encountered "real-world" images in clinics in Thailand that were of lower quality than those in its training set, causing considerable frustration to both patients and staff. Anticipating and mitigating the challenges outlined herein will be key to avoiding such costly failures.

References

1. Verma AA, Murray J, Greiner R, et al. Implementing machine learning in medicine. *CMAJ* 2021;193:E1351-7.
2. Liu Y, et al. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322:1806-16.
3. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. *AJR Am J Roentgenol* 2019;212:513-9.
4. Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med* 2020;26:1229-34.
5. Artificial intelligence and machine learning in software as a medical device. Silver Spring (MD): US Food and Drug Administration; 2021. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (accessed 2020 May 14).
6. Abu-Mostafa YS, Magdon-Ismail M, Lin H-T. *Learning from data: a short course*. AMLbook.com; 2012.
7. Gianfrancesco MA, Tamang S, Yazdany J, et al. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544-7.
8. Cohen JP, Bertin P, Frappier V. Chester: a web delivered locally computed chest x-ray disease prediction system. *ArXiv* 2020 Feb. 2. Available: https://arxiv.org/abs/1901.11210 (accessed 2020 May 14).
9. Cao T, Huang C, Hui DY-T, et al. A benchmark of medical out of distribution detection. uncertainty & robustness in deep learning workshop at ICML. *ArXiv* 2020 Aug 5. Available: https://arxiv.org/abs/2007.04250 (accessed 2020 May 14).
10. Shafaei A, Schmidt M, Little J. A less biased evaluation of out of distribution sample detectors. *ArXiv* 2019 Aug. 20 Available: https://arxiv.org/abs/1809.04729 (accessed 2021 Aug. 16).
11. Ross AS, Hughes MC, Doshi-Velez F. Right for the right reasons: training differentiable models by constraining their explanations. *ArXiv* 2017 May 25. Available: https://arxiv.org/abs/1703.03717 (accessed 2021 Aug. 16).
12. Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018;15:e1002683.

13. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* 2019;2:31.
14. Viviano JD, Simpson B, Dutil F, et al. Saliency is a possible red herring when diagnosing poor generalization. *ArXiv* 2021 Feb. 10. Available: https://arxiv.org/abs/1910.00199 (accessed 2021 Aug. 16).
15. Reed RD, Marks RJ. Neural smithing: supervised learning in feedforward artificial neural networks. Cambridge (MA): MIT Press; 1999.
16. Nestor B, McDermott MBA, Boag W, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *ArXiv* 2019 Aug. 2. Available: https://arxiv.org/abs/1908.00690 (accessed 2021 Aug. 16).
17. Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: *Proceedings of the 32nd International Conference on Machine Learning*; 2015 July 6–11; Lille [FR]. *PMLR* 2015;37:1180-9. Available: http://jmlr.org/proceedings/papers/v37/ganin15.html (accessed 2020 May 14).
18. Bengio Y. Deep learning of representations for unsupervised and transfer learning. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*; 2012 Jun. 28–July 2. Bellevue, Wash. [USA]. *PMLR* 2012;27:17-36. Available: http://proceedings.mlr.press/v27/bengio12a.html (accessed 2020 May 14).
19. Moreno-Torres JG, Raeder T, Alaiz-Rodriguez R, et al. A unifying view on dataset shift in classification. *Pattern Recognit* 2012;45:521-30.
20. Brown MRG, Sidhu GS, Greiner R, et al. ADHD-200 global competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front Syst Neurosci* 2012;6:69.
21. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. *ArXiv* 2014 Apr. 19. Available: https://arxiv.org/abs/1312.6034
22. Cohen JP, Brooks R, En S, et al. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. *ArXiv* 2021 Apr. 24. Available: https://arxiv.org/abs/2102.09475 (accessed 2021 Aug. 16).
23. Powers DMW. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *ArXiv* 2020 Oct. 11. Available: https://arXiv:2010.16061 (accessed 2021 Aug. 16).
24. Seyyed-Kalantari L, Laleh G, McDermott M, et al. CheXclusion: fairness gaps in deep chest x-ray classifiers. *ArXiv* 2020 Oct. 16. Available: https://arxiv.org/abs/2003.00827 (accessed 2021 Aug. 16).
25. Shankar S, Halpern Y, Breck E, et al. No classification without representation: assessing geodiversity issues in open data sets for the developing world. *ArXiv* 2017 Nov. 22. Available: https://arxiv.org/abs/1711.08536 (accessed 2021 Aug. 16).
26. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: *FAT* 2019 — Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*; 2019 Jan. 29–31; Atlanta [GA]. Available: https://doi.org/10.1145/3287560.3287596 (accessed 2020 May 14).
27. Antoniou T, Mamdani M. Evaluation of machine learning solutions in medicine. *CMAJ* 2021 Aug. 30 [Epub ahead of print]. doi:10.1503/cmaj.210036.
28. Blier N. Stories of AI failure and how to avoid similar AI fails [blog]. Amherst (MA): Lexalytics; 2020 Jan. 30. Available: https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2020 (accessed 2021 May 18).
29. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*; 2020 Apr. 25–30; Honolulu. doi:10.1145/3313831.3376718

Competing interests: Joseph Paul Cohen reports receiving CIFAR Artificial Intelligence and COVID-19 Catalyst grants, and a grant from Carestream Health (Stanford University), outside the submitted work. Dr. Cohen also reports that he is Director of the Institute for Reproducible Research, a non-profit organization in the United States. Michael Fralick reports receiving consulting fees from Proof Diagnostics (previously Pine Trees Health), a start-up company developing a CRISPR-based diagnostic test for SARS-CoV-2 infection. No other competing interests were declared.

This article has been peer reviewed.

Affiliations: Mila Quebec AI Institute (Cohen, Viviano, Huang, Bengio), University of Montreal, Montréal, Que.; Vector (Cao, Ghassemi), University of Toronto; Unity Health Toronto (Mamdani); Department of Medicine (Fralick) University of Toronto, Toronto, Ont.; Alberta Machine Intelligence Institute (Greiner), University of Alberta, Edmonton, Alta.; Department of Radiology (Cohen), and Center for Artificial Intelligence in Machine & Imagery (Cohen) Stanford University, Stanford, Calif.

Contributors: All authors contributed to the conception and design of the work, drafted the manuscript, revised it critically for important intellectual content, gave final approval of the version to be published and agreed to be accountable for all aspects of the work.

Content licence: This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY-NC-ND 4.0) licence, which permits use, distribution and reproduction in any medium, provided that the original publication is properly cited, the use is noncommercial (i.e., research or educational use), and no modifications or adaptations are made. See: https://creativecommons.org/licenses/by-nc-nd/4.0/

Correspondence to: Joseph Paul Cohen, joseph@josephpcohen.com