LETTER TO THE EDITOR

Machine learning models and over-fitting considerations

Paris Charilaou, Robert Battat

ORCID number: Paris Charilaou 0000-0002-5512-4225; Robert Battat 0000-0002-7421-9764.

Author contributions: Charilaou P and Battat R drafted and edited the manuscript, and reviewed the intellectual content.

Conflict-of-interest statement: The authors have no conflict of interest to declare.

Country/Territory of origin: United States

Specialty type: Gastroenterology and hepatology

Provenance and peer review: Invited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report’s scientific quality classification
Grade A (Excellent): 0
Grade B (Very good): B, B
Grade C (Good): C
Grade D (Fair): 0
Grade E (Poor): E

Open-Access: This article is an open-access article that was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution NonCommercial (CC BY-NC 4.0) license, which permits others to

©The Author(s) 2022. Published by Baishideng Publishing Group Inc. All rights reserved.

Core Tip: Machine learning models are increasingly being used in clinical medicine to predict outcomes. Proper validation techniques of these models are essential to avoid over-fitting and poor generalization on new data.

Abstract
Machine learning models may outperform traditional statistical regression algorithms for predicting clinical outcomes. Proper validation of building such models and tuning their underlying algorithms is necessary to avoid over-fitting and poor generalizability, which smaller datasets can be more prone to. In an effort to educate readers interested in artificial intelligence and model-building based on machine-learning algorithms, we outline important details on cross-validation techniques that can enhance the performance and generalizability of such models.

Key Words: Machine learning; Over-fitting; Cross-validation; Hyper-parameter tuning

Citation: Charilaou P, Battat R. Machine learning models and over-fitting considerations. World J Gastroenterol 2022; 28(5): 605-607

URL: https://www.wjgnet.com/1007-9327/full/v28/i5/605.htm
DOI: https://dx.doi.org/10.3748/wjg.v28.i5.605

TO THE EDITOR

Con et al[1] explore artificial intelligence (AI) in a classification problem of predicting biochemical remission of Crohn’s disease at 12 mo post-induction with infliximab or adalimumab. They illustrate that, after applying appropriate machine learning (ML) methodologies, ML methods outperform conventional multivariable logistic
regression (a statistical learning algorithm). The area-under-the-curve (AUC) was the chosen performance metric for comparison and cross-validation was performed.

Their study elucidates a few important points regarding the utilization of ML. First, the use of repeated k-fold cross-validation, which is primarily utilized to prevent over-fitting of the models. This technique, while common in ML, has not been traditionally used in conventional regression models in the literature so far. Especially in small datasets, such as in their study \((n = 146)\), linear (and non-linear, in the case of neural networks) relationships risk being “learned” by chance, leading to poor generalization of the models when applied to previously “unseen” or future data points. It was evident from their analysis that the “naïve” AUCs (training the model on all the data), was significantly higher than the mean cross-validated AUCs, in all 3 models, suggestive of “over-fitting” when one does not cross-validate. Smaller datasets tend to be more susceptible to over-fitting as they are less likely to accurately represent the population in question.

Second, the authors utilized “hyper-parameter tuning” for their neural network models, where the otherwise arbitrarily selected “settings” (or hyper-parameters, such as the number of inner neuron layers and number of neurons per layer) of the neural network are chosen based on performance. Hyper-parameters cannot be “learned” or “optimized” by simply fitting the model (as it happens with predictor coefficients), and the only way to discover the best values is by fitting the model with various combinations and assessing its performance. The combinations can be evaluated stochastically (randomly or via a Bayes-based approach) or using a grid approach (e.g., for 3 hyper-parameters that take 5 potential values, there are \(5 \times 5 \times 5 = 5^3 = 125\) combinations to evaluate) over \(k\) times. One may ask, if one was to fit a model \(125 \times k\) times, on 146 observations, is there not a risk for over-fitting the “optimal” hyper-parameter values? To avoid such a problem, nested k-fold cross-validation must be performed: within each repeated k-fold training data subset, a sub-k-fold “inner” training/validation must be done to evaluate each hyper-parameter combination. In this way, we overcome potential bias to optimistic model performance, which can occur when we use the same cross-validation procedure and dataset to both tune the hyper-parameters and evaluate the model’s performance metrics (e.g., AUC)[2]. The authors did not elaborate on how the hyperparameter tuning was performed.

Another point to consider in k-fold cross-validation in small datasets is the number of k-folds used, specifically in classification problems (i.e., yes/no binary outcomes). In this study[1], the outcome prevalence was 64\% \((n = 93)\). With a chosen \(k = 5\), the training folds would comprise 80\% of that data, leading to approximately 74 positive cases of biochemical remission. The number of positive outcomes in each training fold must be considered, especially in logistic regression, where the rule of thumb recommends at least ten positive events per independent predictor, to minimize over-fitting[3]. In this study[1], six predictors were eventually used in the multivariable model, making over-fitting less likely from a model-specification standpoint. Finally, k-folds are recommended to be stratified by the outcome, so the outcome prevalence is equal among the training and testing folds. This becomes crucial when the prevalence of outcome of interest is \(< 10\%-20\\%\) (imbalanced classification problem). While imbalanced classification is not an issue in this study[1], the authors did not mention whether they used outcome-stratified k-folds.

Lastly, the endpoint utilized, CRP normalization, has poor specificity for endoscopic inflammation in Crohn’s disease[4]. More robust endpoints would include endoscopic inflammation and/or deep remission using validated disease activity indices[5].

We congratulate the authors for their effort, which acts both as a proof-of-concept for using ML in improved prediction of outcomes in IBD, but also for the methodologies outlined to reduce over-fitting. In general, with the advent of AI and specifically ML-based models in IBD[6], it is important to recognize that while now we have the tools to construct more accurate models and enhance precision medicine, most ML-based models, such as artificial neural networks, lack in being intuitively interpretable (i.e., “black-box”). Efforts in “explainable AI” are under way[7], hopefully eliminating the “black-box” concept in future clinical decision tools. Applying these to validated disease activity assessments will be essential for prediction models in future studies.

REFERENCES

1 Con D, van Langenberg DR, Vasudevan A. Deep learning vs conventional learning algorithms for clinical prediction in Crohn’s disease: A proof-of-concept study. World J Gastroenterol 2021; 27: 6476-6488 [PMID: 34720536 DOI: 10.3748/wjg.v27.i38.6476]
2 Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010; **11**: 2079-2107

3 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49**: 1373-1379 [PMID: 8970487 DOI: 10.1016/s0895-4356(96)00236-3]

4 Mosli MH, Zou G, Garg SK, Feagan SG, MacDonald JK, Chande N, Sandborn WJ, Feagan BG. C-Reactive Protein, Fecal Calprotectin, and Stool Lactoferrin for Detection of Endoscopic Activity in Symptomatic Inflammatory Bowel Disease Patients: A Systematic Review and Meta-Analysis. *Am J Gastroenterol* 2015; **110**: 802-19; quiz 820 [PMID: 25964225 DOI: 10.1038/ajg.2015.120]

5 Turner D, Ricciuto A, Lewis A, D’Amico F, Dhaliwal J, Griffiths AM, Bettenworth D, Sandborn WJ, Sands BE, Reinisch W, Schölmerich J, Bemelman W, Danese S, Mary JY, Rubin D, Colombel JF, Peyrin-Biroulet L, Dotan I, Abreu MT, Dignass A; International Organization for the Study of IBD. STRIDE-II: An Update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) Initiative of the International Organization for the Study of IBD (IOIBD): Determining Therapeutic Goals for Treat-to-Target strategies in IBD. *Gastroenterology* 2021; **160**: 1570-1583 [PMID: 33359090 DOI: 10.1053/j.gastro.2020.12.031]

6 Chen G, Shen J. Artificial Intelligence Enhances Studies on Inflammatory Bowel Disease. *Front Bioeng Biotechnol* 2021; **9**: 635764 [PMID: 34307315 DOI: 10.3389/fbioe.2021.635764]

7 Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel)* 2020; **23** [PMID: 33375658 DOI: 10.3390/e23010018]
