Concerns regarding "*Development of a machine learning-based model to predict hepatic inflammation in chronic hepatitis B patients with concurrent hepatic steatosis: a cohort study*"

Zhi Geng,[a,b,c,e] Tao Guo,[d,e] Ling Wei,[a,b,c,*] and Kai Wang[a,b,c,**]

[a]Department of Neurology, The First Affiliated Hospital of Anhui Medical University, Hefei, China
[b]Anhui Province Key Laboratory of Cognition and Neuropsychiatric Disorders, Hefei, China
[c]Collaborative Innovation Centre of Neuropsychiatric Disorder and Mental Health, Hefei, China
[d]Center for Biomedical Imaging, University of Science and Technology of China, Hefei, Anhui, China

eClinicalMedicine
2024;78: 102907
Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102907

I recently read the article published in your journal, titled "Development of a machine learning-based model to predict hepatic inflammation in chronic hepatitis B patients with concurrent hepatic steatosis: a cohort study".[1] In this study, the authors propose that a Gradient Boosting Classifier (GBC) machine learning model constructed on the basis of simple, common clinical indicators can accurately predict moderate-to-severe hepatitis status in patients with chronic hepatitis B combined with fatty liver disease. However, after a careful reading of the original article and the supplementary material, I identified several issues in the methodology section that may have affected the study's conclusions. We would like to make relevant comments and suggestions to the editorial board and the authors to further enhance the scientific validity, reliability, and readability of this study.

Controversy over failure to split the training cohort dataset

In this study, we did not observe in the methodology or elsewhere in the paper that the authors split the entire training cohort dataset, i.e., the process of splitting it into a certain percentage training set data for training and other percentage test set for testing. In the field of machine learning, this is not the norm or even this is a very rare practice.[2,3] Even though the authors used data from two independent external validation centers for validation, these data were only used to test the generalization ability of the machine learning model.[2] If the authors did not split the entire training cohort dataset into separate sets for training and testing, it raises another issue. Specifically, in the results presented in Fig. 3A of this paper, are the individual model ROCs obtained from the training cohort derived from the entire training cohort dataset? Using the entire training cohort dataset to both train the model and test it is clearly incorrect, rendering the conclusions invalid. This practice can lead to risks such as data leakage, distorted model performance metrics, and model overfitting.

If the authors are splitting the entire training cohort dataset into five parts for five-fold cross-validation, this introduces another problem. Specifically, the individual model ROC curves shown in Fig. 3A for the training cohort are based on the weighted performance values of each validation fold. The legend in the ROC curves should indicate the mean and standard deviation of the ROC. The authors do not specify which part of the training cohort dataset the ROC graph in Fig. 3A is based on.

The two independent external validation datasets are used solely to test the generalization ability of the model, which is entirely different from partitioning the entire dataset for training and internal testing of the model. It would be helpful to readers if the authors provided further clarification on this point.

Controversy over variable screening

In this study, SHAP (SHapley Additive exPlanations) method-based variable importance is used for variable screening. The importance of variables identified by the SHAP method primarily depends on the parameters of the machine learning model and the data used for training.[4] In this study, the authors did not specify whether each machine learning model used default parameters or performed a hyper-parameter optimization search. This distinction is crucial as it would have produced different variable importance values, further affecting the selection of variables and the final conclusions. Additionally, the authors did not clarify whether the variable importance was generated using the overall dataset of the training cohort or by dividing

DOI of original article: https://doi.org/10.1016/j.eclinm.2023.102419
*Corresponding author. Department of Neurology, The First Affiliated Hospital of Anhui Medical University, Hefei, China.
**Corresponding author. Department of Neurology, The First Affiliated Hospital of Anhui Medical University, Hefei, China.
E-mail addresses: ahykdxwl@126.com (L. Wei), wangkai1964@126.com (K. Wang).

[e]These authors contributed to the work equally and should be regarded as co-first authors.

the training cohort dataset. If the former is the case, it can lead to data leakage, resulting in overfitting or overestimation of the model, which ultimately affects the validity of the research conclusions. If the entire training cohort dataset is divided using five-fold cross-validation, there would be a SHAP importance ranking for each of the five trials. Is the ranking of feature importance shown in Fig. 2 an average of the importance scores from these five experiments? The authors should provide a detailed explanation of the origin of the SHAP importance values.

Although the authors adopted an approach of continuously accumulating and iterating based on the variables to obtain the combination that maximizes the ROC of the model, which is an effective method for variable selection. It is not specified in this study whether this iterative algorithm was performed on the overall training cohort dataset or after the dataset was divided. This distinction is significant for the reliability of the results and is not addressed in the study. If the study was trained and tested without further segmentation of the overall training cohort dataset, it significantly increases the risk of data leakage, which is not a scientific and reliable practice.

Additionally, the iterative approach to screening by variable importance introduces the problem of obtaining different combinations of variables for each model. This lack of consistency makes multi-model comparisons incomparable because the benchmarks for these comparisons are not uniform. Even if a model performs well, it is difficult to determine whether its performance is due to the model's inherent strength or the choice of variables, creating a dilemma for interpreting the results.

Controversy about hyperparameter optimization of various machine learning models

In this study, the process of hyperparameter optimization is not described, even though different combinations of hyperparameters are crucial for model building and validation.

In the field of machine learning, using default hyperparameters may not be suitable for specific datasets. Lack of hyperparameter optimization can prevent the model from fully exploiting feature information, leading to less stable performance, high heterogeneity, and insufficient robustness, which negatively affects model building and generalization. Different combinations of hyperparameters significantly impact the predictive efficacy of the model. Failure to adequately perform hyperparameter optimization may result in models that do not achieve optimal predictive efficacy. This causes selection bias in the model and ultimately leads to incorrect decisions.

Machine learning is prone to model overfitting and poor generalization if the built-in hyperparameters are not properly set.[5] Therefore, hyperparameter optimization in the field of machine learning is a scientific, rigorous, and reliable practice.

The rationale for selecting the GBC machine learning model as the optimal choice for building the web calculator is not fully explained in this study

The rationale for selecting the GBC machine learning model as the best model for building the web calculator is not fully explained in this study. The authors do not explicitly describe how the training cohort data is divided in the model development section. In the case of a five-fold cross-validation division, the final result is a GBC model with five different parameter sets. These should be considered when developing the web calculator to specify a particular GBC model. It is not stated whether the model used is one of the five GBC models from the five-fold cross-validation or if it was redeveloped based on the entire training cohort. This ambiguity directly affects the correctness of the study's conclusions. Clarifications or corrections from the authors would be appreciated.

Contributors

ZG, TG: performed the data analysis, drafted original manuscript, revised the draft paper, critical revision of the manuscript for important intellectual content. LW, KW: designed, conceived and supervised the study. All the authors have read and approved the final manuscript.

Declaration of interests

We declare no competing interests.

References

1 Rui F, Yeo YH, Xu L, et al. Development of a machine learning-based model to predict hepatic inflammation in chronic hepatitis B patients with concurrent hepatic steatosis: a cohort study. *eClinicalMedicine*. 2024;68:102419.
2 Zou H, Lu Z, Weng W, et al. Diagnosis of neurosyphilis in HIV-negative patients with syphilis: development, validation, and clinical utility of a suite of machine learning models. *EClinicalMedicine*. 2023;62:102080.
3 Geng Z, Yang C, Zhao Z, et al. Development and validation of a machine learning-based predictive model for assessing the 90-day prognostic outcome of patients with spontaneous intracerebral hemorrhage. *J Transl Med*. 2024;22(1):236.
4 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–4774.
5 Forrest IS, Petrazzini BO, Duffy A, et al. Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. *Lancet*. 2023;401(10372):215–225.