PLOS Computational Biology

FORMAL COMMENT

Comment on "Using genomic data and machine learning to predict antibiotic resistance: A tutorial paper"

Davide Chicco[1,2*], **Giuseppe Jurman**[3,4]

1 Dipartimento di Informatica Sistemistica e Comunicazione, Università di Milano-Bicocca, Milan, Italy, **2** Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada, **3** Department of Biomedical Sciences, Humanitas University, Milan, Italy, **4** Data Science for Health Unit, Fondazione Bruno Kessler, Trento, Italy

* davidechicco@davidechicco.it

OPEN ACCESS

Citation: Chicco D, Jurman G (2025) Comment on "Using genomic data and machine learning to predict antibiotic resistance: A tutorial paper". PLoS Comput Biol 21(12): e1013673. https://doi.org/10.1371/journal.pcbi.1013673

Editor: Patricia M Palagi, SIB: Swiss Institute of Bioinformatics, SWITZERLAND

Received: January 15, 2025

Accepted: October 28, 2025

Published: December 1, 2025

Funding: The work of D.C. is partially funded by the Italian Ministero Italiano delle Imprese e del Made in Italy under the Digital Intervention in Psychiatric and Psychologist Services (DIPPS) programme and is partially supported by Ministero dell'Università e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023-2027" ReGAInS grant assigned to Dipartimento

Abstract

A recent study by Faye Orcales and colleagues proposes a teaching curriculum on supervised machine learning applied to genomics data aimed at predicting antibiotic resistance. The article describes a traditional machine learning pipeline step-by-step in a way that is accessible to anyone, including novices. However, the authors provide a misleading piece of advice in the "Evaluating model performance" section, where they recommend that readers use accuracy and the F1 score for binary classification. We write this short formal comment on that article to reaffirm and explain why accuracy and the F1 score should be avoided in the evaluation of binary classification and why the Matthews correlation coefficient (MCC) should be employed instead. We also take this opportunity to warn readers about the dangers of *k*-fold cross-validation, which is suggested as a standard method for dividing data into training set and test set, but has several flaws and pitfalls.

Formal comment

A recent study published by Faye Orcales et al. [1] in PLOS Computational Biology thoroughly describes a teaching proposal for a tutorial on supervised machine learning applied to genomics data for predicting antibiotic resistance. Antibiotic resistance is a significant global public health problem indeed, and computational intelligence can be an effective tool for analyzing genomics data and for identifying potential interesting data trends related to antibiotic resistance.

The article effectively outlines the datasets used, the machine learning models employed, common practices in machine learning such as cross validation, and various binary classification evaluation metrics. It also provides an overview of the contents of the six Google Colab notebooks made available by the authors to the students participating in the tutorial.

di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca.The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

In the "Evaluating model performance" section, the authors recommend students and readers of the article calculate four metrics: accuracy, recall (true positive rate), precision (positive predictive value), and F1 score.

While we agree on the usage of recall and precision, we completely disagree with the authors regarding the F1 score and accuracy.

As we have explained in the past [2], F1 score and accuracy can be misleading when handling unbalanced datasets and, therefore, should be avoided. Instead, the Matthews correlation coefficient (MCC) should be emphasized as the most valuable metric for binary classification evaluations.

Let us suppose we have a dataset of ten elements, consisting of nine positives and one negative, and a naive classifier that always predicts 1s. The predicted values would be {1, 1, 1, 1, 1, 1, 1, 1, 1, 1}.

In this case, the accuracy would be 0.900, and the F1 score would be 0.947 (on a scale from 0 to 1), which could be interpreted as almost perfect prediction. However, if we examine the confusion matrix, we would find that we have TP (True Positives) = 9, FP (False Positives) = 1, and no TN (True Negatives) and no FN (False Negatives), indicating low quality.

The MCC, on the other hand, would be undefined in this scenario, raising a red flag for the researcher or student.

Several other similar examples could be mentioned. Consider a case where we have TP = 90, FP = 5, TN = 1, and FN = 4. Even if this classifier correctly predicts all 95 positive elements, it clearly performs poorly in predicting the negatives: only one out of five negatives is correctly identified as a true negative.

In this imbalanced case, a wise metric would yield a low result. However, accuracy would be 0.91, and the F1 score would be 0.9524 (on a scale from 0 to 1), falsely suggesting an excellent prediction. The MCC, in contrast, would be 0.135 (on a scale from -1 to +1), indicating a result similar to random chance.

Surprisingly, the authors of the [1] study are aware of this trouble with accuracy (they wrote: "For example, if 95% of the isolates in the data set are susceptible, and a model predicts that all isolates are susceptible, then the accuracy would be 95%, because the model gets it right for 95% of the cases; 95% accuracy sounds great"), but they suggest to handle it by using recall (true positive rate) together with accuracy.

We disagree and recommend anyone performing a binary classification to calculate the MCC and the four confusion matrix basic rates, such as recall (true positive rate), precision (positive predictive value) but also specificity (true negative rate) and negative predictive value.

To better explain this point, we propose the example of isolate classification including the outcome measured through the Matthews correlation coefficient in Fig 1.

In that example, we have five bacterial isolates that were predicted to be resistant or susceptible to an antibiotic. We compared these five predicted labels with the original antibiotic labels of the five isolates to generate a confusion matrix having 3 TP, 1 FN, 1 FP, and no TN.

A. Prediction results

sample	actual label	predicted label	result
isolate 1	resistant to the antibiotic	resistant to the antibiotic	TP (true positive)
isolate 2	resistant to the antibiotic	resistant to the antibiotic	TP (true positive)
isolate 3	resistant to the antibiotic	susceptible to the antibiotic	FN (false negative)
isolate 4	resistant to the antibiotic	resistant to the antibiotic	TP (true positive)
isolate 5	susceptible to the antibiotic	resistant to the antibiotic	FP (false positive)

B. Confusion matrix

(actual labels) \ (predicted labels)	resistant	susceptible
resistant	TP: 3	FN: 1
susceptible	FP: 1	TN: 0

C. Performance metrics

accuracy = 0.60 in the [0,1] interval
F1 score = 0.75 in the [0,1] interval
MCC = −0.25 in the [−1,+1] interval

Fig 1. Example of how confusion matrix results are used to calculate evaluation metrics. (A) Determination of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) labels. (B) The confusion matrix based on the (A) model showing total numbers of TP, TN, FP, and FN. (C) Numbers from the confusion matrix are used to calculate various evaluation metrics.

https://doi.org/10.1371/journal.pcbi.1013673.g001

We then calculated accuracy, F1 score, and MCC for this confusion matrix, and the MCC resulted being the only informative truthful outcome of this experiment (Fig 1)

As one can notice, accuracy and F1 score produced high results, indicating overoptimistic 60% to 75% levels of correctness, respectively. The MCC, instead, produced a low negative value, clearly indicating a problem in this binary classification. The problem, in fact, is the lack of true negatives (TN). Accuracy and F1 score were unable to spot this drawback.

Even if the Matthews correlation coefficient has several advantages, we need to acknowledge that it is not a silver bullet and has some flaws, too: in particular, it can be undefined when two classes of the confusion matrix are zero [3].

Another flaw in the study is the recommendation to use cross-validation. While it is a common practice, it has several disadvantages, as the assignment of data to the *k*-folds is always arbitrary and represents only one of the possible partitions of the dataset [4].

Instead of using cross-validation, we recommend repeated hold-out validation: splitting the dataset into 80% randomly selected data elements for the training set, using the remainder for the test set, calculating the results, and then repeating this process one thousand times to report the average results. Of course, the training set /test set partition should be stratified, to keep the same general ratio of positives and negatives of the whole dataset. This approach ensures that at each iteration, a random subsample of the data is selected for the training set, making the final results more universal and generalizable than those obtained from a single short partition made through cross-validation.

The article by Faye Orcales et al. [1] also merits recognition for not mentioning and not promoting the receiver operating characteristic (ROC) curve and its area under the curve (AUC), which is a misleading index still commonly employed in machine learning studies, unfortunately [5–7].

We firmly believe that the ROC AUC should be avoided whenever possible. When ROC AUC is needed to be reported for comparison with other studies, it should be discussed thoroughly with care, highlighting its critical points.

In conclusion, we find most of the [1] article interesting and useful for students who want to learn machine learning applied to antibiotic resistance data. However, we invite the authors to advise readers and students to use the Matthews

correlation coefficient to assess binary classification results, and to stay away from accuracy and F1 score in any future tutorials they might release. Moreover, we invite them to illustrate the advantages of repeated hold-out validation compared to cross-validation in their tutorial.

Author contributions

Conceptualization: Giuseppe Jurman.

Formal analysis: Davide Chicco.

Investigation: Davide Chicco, Giuseppe Jurman.

Methodology: Davide Chicco, Giuseppe Jurman.

Project administration: Davide Chicco.

Supervision: Davide Chicco, Giuseppe Jurman.

Validation: Giuseppe Jurman.

Writing – original draft: Davide Chicco.

Writing – review & editing: Davide Chicco, Giuseppe Jurman.

References

1. Orcales F, Moctezuma Tan L, Johnson-Hagler M, Suntay JM, Ali J, Recto K, et al. Using genomic data and machine learning to predict antibiotic resistance: A tutorial paper. PLoS Comput Biol. 2024;20(12):e1012579. https://doi.org/10.1371/journal.pcbi.1012579 PMID: 39775233
2. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:6. https://doi.org/10.1186/s12864-019-6413-7
3. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 2021;14(1):13. https://doi.org/10.1186/s13040-021-00244-z PMID: 33541410
4. Rao RB, Fung G, Rosales R. "On the dangers of cross-validation. An experimental evaluation." Proceedings of the 2008 SIAM international conference on data mining. Society for Industrial and Applied Mathematics. 2008. https://doi.org/10.1137/1.9781611972788.54
5. Muschelli J. ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. J Classif. 2020;37(3):696–708. https://doi.org/10.1007/s00357-019-09345-1 PMID: 33250548
6. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography. 2007;17(2):145–51. https://doi.org/10.1111/j.1466-8238.2007.00358.x
7. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining. 2023;16:4. https://doi.org/10.1186/s13040-023-00322-4