American Journal of Emergency Medicine xxx (xxxx) xxx

Contents lists available at ScienceDirect

American Journal of Emergency Medicine

journal homepage: www.elsevier.com/locate/ajem

Reassessing predictive modeling for emergency department return in COVID-19 patients

Fong et al. developed a machine learning model to predict emergency department (ED) return for COVID-19 patients using Extreme Gradient Boosting (XGBoost) trained on data from two regional health systems [1]. By analyzing 42,056 encounters, the study highlighted the potential of machine learning to support ED disposition decisions and optimize discharge planning. Feature importance was evaluated using XGBoost with SHAP and Gini metrics to identify predictors, including age, vital signs, and specific laboratory markers. Despite promising results, several considerations about the methodology warrant discussion [1]. While cross-validation using metrics such as AUC, RMSE, and R-squared effectively evaluates the accuracy of target predictions, it falls short in assessing the accuracy of feature importance due to the lack of ground truth values.

While machine learning approaches like XGBoost are powerful, the reliability of feature importance measures such as SHAP and Gini values remains a critical concern, leading to potentially erroneous conclusions [2,3]. These metrics are inherently model-dependent and can propagate non-negligible biases present in the underlying data. For example, while SHAP values are beneficial for interpreting individualized predictions, they inherently assume that features are independent. This assumption often fails in datasets with correlated predictors, as is frequently the case with clinical data [4]. Similarly, Gini values, reflecting overall model performance, may exaggerate the significance of features influenced by overfitting, particularly in models trained on imbalanced or incomplete datasets [5].

Understanding the biases in feature importance measures for XGBoost and SHAP requires an in-depth exploration of their methodologies and assumptions. XGBoost, an ensemble learning algorithm based on decision trees, builds models sequentially to reduce errors from earlier iterations. Its methods for assessing feature importance can be biased due to the use of metrics like gain, which reflects a feature's contribution to accuracy improvements; cover, indicating the proportion of observations associated with a feature; and frequency, the number of times a feature is used across trees. These metrics are model-dependent, meaning different configurations or parameters can yield varying importance rankings, which are not absolute indicators of predictive relevance.

Feature correlations add complexity. Correlated features can distort importance scores as the model arbitrarily distributes "credit," overstating or understating individual contributions. Overfitting is another concern, with XGBoost potentially inflating importance scores for features specific to the training data but irrelevant for new predictions. Additionally, the sequential tree-building process can disproportionately highlight features that correct earlier errors, exaggerating their importance.

SHAP, while powerful for interpreting feature importance, also has limitations. SHAP values depend on the training data and model structure, which can amplify inherent model biases. SHAP assumes feature independence when estimating contributions, an assumption often invalid in real-world scenarios, especially when features are correlated. This correlation can distort SHAP values, either masking or overstating individual feature effects. Outliers or noise further complicate SHAP's reliability, as they may disproportionately influence importance measures. Moreover, SHAP struggles with non-linear relationships in complex models, potentially misrepresenting feature contributions.

Both XGBoost and SHAP, despite being influential tools, introduce biases through their dependencies and assumptions. Researchers should complement machine-learning-derived importance metrics with rigorous statistical analyses to achieve a clearer understanding of feature relationships and predictive relevance.

Fong et al.'s reliance on metrics conditioned by model structure raises questions about the generalizability of their findings. The variations in model performance across health systems—with HS2 data consistently yielding better metrics—highlight the influence of regional differences and data quality. To enhance robustness, future studies should consider additional statistical techniques, such as Spearman's correlation with *p*-values and Kendall's tau with p-values, along with non-linear and non-parametric approaches to validate associations between the target and features independently of the model [6,7]. These methods offer a bias-free complement to machine learning metrics and could uncover genuine relationships.

Another critical issue is the model's prioritization of recall over precision, reflecting a focus on sensitivity at the expense of specificity. While this approach supports identifying high-risk patients, the resulting trade-off in precision suggests a potential for false positives, which may strain ED resources. Enhancing the model's ability to balance these metrics, perhaps by incorporating nuanced data such as social determinants of health or unstructured text from clinician notes, could significantly improve its utility as a clinical decision support tool [8].

Finally, the study's discussion of predictive features highlights the potential for clinical misinterpretation. For example, the finding that higher SpO2 levels reduce the likelihood of ED return aligns with clinical expectations, but the inclusion of certain predictors, such as past use of 5HT3 receptor antagonists, requires careful contextualization to avoid overgeneralization [9]. Ensuring that feature importance metrics align with clinical plausibility is crucial to maintaining the trust and utility of machine learning tools in healthcare.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

https://doi.org/10.1016/j.ajem.2025.01.009

Please cite this article as: N.M. Hieu and Y. Takefuji, Reassessing predictive modeling for emergency department return in COVID-19 patients, American Journal of Emergency Medicine, https://doi.org/10.1016/j.ajem.2025.01.009

Consent for publication

Not applicable.

CRediT authorship contribution statement

Nguyen Minh Hieu: Conceptualization, Investigation, Validation, Writing – original draft, Writing – review & editing. **Yoshiyasu Takefuji:** Conceptualization, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare no conflicts of interest.

References

[1] Fong A, Adams KT, Khairat S, Galarraga JE. Using machine learning to predict emergency department return across two regional health systems; a generalizable model for COVID-19 patients. Am J Emerg Med. 2024. https://doi.org/10.1016/j.ajem.2024.11.032.
[2] Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20:177.
[3] Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711–8. https://doi.org/10.1093/bioinformatics/bty373.
[4] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785–94. https://doi.org/10.48550/arXiv.1603.02754.
[5] Adler Afek, Painsky Amichai. Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection; 2021. https://doi.org/10.48550/arXiv.2109.05468.
[6] Tyagi A, Salhotra R, Agrawal A, Vashist I, Malhotra RK. Use of Pearson and Spearman correlation testing in Indian anesthesia journals: an audit. J Anaesthesiol Clin Pharmacol. 2023;39(4):550–6. https://doi.org/10.4103/joacp.joacp_13_22.
[7] Jiang Jiefang, Zhang Xianyong, Yuan Zhong. Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets. Exp Syst Appl. 2024;249:123633. https://doi.org/10.1016/j.eswa.2024.123633.
[8] Trzeciak S, Rivers EP. Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. Emerg Med J. 2003;20(5):402–5. https://doi.org/10.1136/emj.20.5.402.
[9] Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China [published correction appears in lancet. 2020 Feb 15;395(10223):496. doi: 10.1016/S0140-6736(20)30252-X]. Lancet. 2020;395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

Nguyen Minh Hieu
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
E-mail address: g2450010@stu.musashino-u.ac.jp

Yoshiyasu Takefuji
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
*Corresponding author.
E-mail address: takefuji@keio.jp

19 December 2024
Available online xxxx