Peer Review File

Article information: https://dx.doi.org/10.21037/jhmhp-23-97

Reviewer A	Response
1. My main concern is the title. The title says they did a comparison of machine learning algorithms, which is true, but if that was the main purpose of the study, I would expect the authors to use more current machine learning algorithms like XGBoost etc. The authors use some more classical machine learning algorithms, which might be considered by the modern data science as rather outdated. So I don't think that the title really reflects well was done in the study. My take on the study is that they want to develop a predictor calculator of intracranial hemorrhage in the elderly population. If that was indeed the main aim, then they need to change the title and I think that the methodology would then be appropriate. Again, I can argue that they used less modern machine learning algorithms, but there is nothing wrong with that.	-I changed the title to "Web-Based Calculator Using Machine Learning to Predict Intracranial Hematoma in Geriatric Traumatic Brain Injury" (Change in Title) - I added analysis of extreme gradient boosting (XGB) algorithm in the revised manuscript. (Change in Result section, Table 3, Figure 3 and Figure 4)
2. I would like to highlight two omissions from the methodology. One is that they didn't calibrate the model and second that they didn't explicitly follow some of the established guidelines like the TRIPOD guidelines. These should either be addressed or listed as limitations.	- The revised manuscript followed the TRIPOD guidelines, which included sample size, missing data management, and calibration - Calibrate model by the Hosmer-Lemeshow test (goodness of fit test) (Change in text:Line 143-145, 199-200) - Sample size (Change in text:line 106-109) - Missing data management (Change in text:line 136-137)
3. Their methodology of feature selection where they basically used techniques from logistic regression is not invalid, but it's rather unconventional, given that they are dedicated machine learning methodologies for feature selection. My concern with their approach is that they might have missed some non-linear relationships between their predictor variables by using logistic regression-based feature selection. That might weaken the performance of their model, but I don't see that leading to any methodological erroneous results.	- I changed the methodology of feature selection from logistic regression with a backward stepwise procedure to a chi-square test for categorical variables and an independent t-test for continuous variables to avoid missing some non-linear relationships between their predictor variables. Therefore, we re-analyzed 20 predictors and reported new results in the revised manuscript. (Change in text:Line140-144, 194-200) - I revised table 2 (feature selection by chi-square test and t-test)

4. Overall, I find it as a good effort and I like the fact that they made the calculator public available.	- Web application was updated according to revised manuscript.
5. One final note is that they need to provide some more information about the patient population. Was that a trauma center? Urban or rural? Etc	- Urban trauma center (Change in text:Line 103-014)
Reviewer B	
In this manuscript, the authors investigate the potential of machine learning algorithms, with a focus on the Naïve Bayes (NB) algorithm, to predict traumatic intracranial hematomas in elderly traumatic brain injury (TBI) patients. The study addresses a pertinent issue, as elderly individuals are at a higher risk of developing cerebral hematomas following TBI, leading to the overuse of cranial computed tomography (CT). Overall, the paper shows promise but requires some improvements for publication. Strengths: Relevance and Significance: The study addresses an important issue by exploring the use of ML algorithms in predicting intracranial hematomas, which could potentially reduce unnecessary CT scans among elderly TBI patients. Methodological Approach: The use of a retrospective cohort study and the application of various ML algorithms, including NB, provide a robust foundation for the research. The separation of the dataset into training and testing sets is appropriate for model evaluation. Results: The presentation of results is clear and concise, highlighting the NB algorithm's high sensitivity and acceptable AUC. Areas for Improvement: Clarity and Organization: The abstract could benefit from a more structured presentation. It should clearly state the research problem, objectives, methods, results, and conclusions. Additionally, the abstract lacks key details such as the number of patients	- I revised the abstract by adding the number of patients with positive CT scans (Change in Table 1, Abstract line 40, and results) - I added papers to the literature review according to reviewer's suggestion. (Change in text:line 244-253)

with positive CT scans, making it less informative. I also would like to know if there are any overlapping subjects between the train and test sets. Contextualization: The paper needs more context regarding the existing literature on this topic. It would be beneficial to include more reviews of relevant studies or techniques in the field of TBI and ML to highlight the novelty or significance of the research.	
In conclusion, this manuscript shows potential, but it needs revision to improve its clarity, organization, and contextualization within the broader field. Additionally, consider expanding on the discussion of the practical implications of the findings.	- I revised conclusion in the present study according to reviewer's suggestion. (Change in text:Line 293-297)
Reviewer C	
1. According to the paper, the proposed method will be used as a screening tool for hematoma prediction, where recall and precision are important. I suggest also reporting AUPRC.	- I added AUPRC in the present study according to reviewer's suggestion (Change in text:Line 161-166, Figure 4, Table 3).
2. The dataset is quite unbalanced. Did the authors try any strategy to avoid the bias from unbalanced data?	- Unbalanced numbers in clinical outcomes are a common problem in medicine, therefore, we added AUPRC and I also reported an F1-score for unbalanced outcomes in the present study. (Change in text:Line 161-166, Figure 4, Table 3)
3. It would be great to add more discussion on what level of performance is acceptable for the model to be used in practice.	- I added further discussion about level of performance according to reviewer's suggestion. (Change in text:Line 209-216)
4. More analysis of the model's predictions will be constructive to understand the model's performance better. For example, as the dataset contains annotations for different types of intracranial hematoma (e.g., EDH, SAH), it might be interesting to evaluate the distribution of different types of intracranial hematoma in misclassified cases and correctly predicted cases.	In real-world clinical practice, I admitted all patients who add any types of intracranial hematoma that caused I grouped all type of intracranial hematoma. Additionally, I will conduct another paper for prognostic factor of elderly traumatic brain and various types of intracranial hematoma is one of predictors of their prognosis.
5. The ROC curves in Figure 3 are a bit strange. Why does each curve only have three points (usually the curve is drawn	- If ROC curves were developed by binary classifier prediction, figure of ROC curve will be as prior Figure 3. However, ROC

from a number of points, i.e., multiple levels of specificity)?	curves in the revised manuscript were developed from predictive probabilities (continuous variable) and ROC curves look more smooth curve as updated Figure 3.
6. Table 4 should be Table 3	- I edited it according to reviewer's suggestion (Change in Table3)