Peer Review File

Article information: https://dx.doi.org/10.21037/jtd-24-711

Reviewer A

The authors have compared the efficacy of eight different machine learning algorithms to assess the risk of AKI triggered by OPCAB. They concluded that the gradient-boosting decision tree (GBDT) is the most superior. This study is intriguing and of significant importance. However, I have the following concerns that I believe should be addressed in a revised manuscript:

Comments 1: Primarily, this study appears more as research in artificial intelligence rather than in cardiovascular surgery, and it is unclear whether it is appropriate for publication in the Journal of Thoracic Disease.

Reply 1: First of all, the research population of this study is OPCABG patients, the main purpose is to identify high-risk patients with OPCABG-AKI early through the model, and improve the prognosis of OPCABG patients. Besides,different machine learning algorithms are the means to predict OPCABG-AKI, and the best prediction model is selected by comparing different models, which is also to better guide perioperative clinical decision-making of OPCABG. In conclusion, this study is related to cardiovascular surgery.

Changes in the text: None.

Comments 2: The paper introduces several machine learning models for predicting AKI but lacks detailed justification for the selection of these models and how their parameters were optimized during the study. A more comprehensive explanation of the reasons behind choosing these specific models and the methods used for parameter tuning would enhance the manuscript.

Reply 2: Reasons for selecting the machine learning model: based on previous studies, namely references [14]-[17], and potential linear and nonlinear data relationships. Parameter adjustment method: The length of the preprocessed intraoperative dataset L was 18,000, and the number of feature data N was 28. During the intraoperative data segmentation stage, K was set to 180, P was set to 120, zero padding length was 90, final S was 150, and the 3-D tensor was obtained after segmentation combination.Each deep learning model was pre-trained for 100 epochs on an 18,000-second sequence, and the training was terminated early if the best model was not found in the validation set for 30 consecutive epochs. The initial learning rate was 1e-3, and the decay rate was 0·97 every two epochs. The optimizer used was Adam30, and the pre-trainingloss function was the traditional Cross-Entropy Loss.Accuracy and area under the curve (AUC) 31 were calculated for each model (both machine learning models and deep learning models), with AUC serving as the primary evaluation metric for determining the model's performance. In order to address the issue of an unbalanced class, the loss function's weight was also added to the models.

Changes in the text: None.

Comment 3: It would be beneficial to include more details about the dataset, such as the characteristics of the patient population, the handling of missing data, and the cross-validation process used in the machine learning analysis. These details will improve the reproducibility and credibility of the study.

Reply 3:

①The characteristics of the patient population：

This study is not about simple component comparison, but focuses on machine learning model prediction. There is no problem of component comparison in this algorithm, and only two overall data sets, training set and test set, exist. The specific cases in the training and test sets were randomized for both the AKI and non-AKI groups. And the machine learning operation is constantly re-randomizing the combination of cases, each time the two groups of patients change, for example, case 1 in the first operation in the training set, maybe the second is randomized into the test set. Therefore, the case selection of machine learning is made by the machine randomly, each time is not repeated, the position of the cases is not fixed, and the cases will not be selected by humans to enter the group. Therefore, the baseline characteristics cannot be provided.

②Handling of missing data:

Preoperative data:Due to the fact that some of the preoperative data obtained at follow-up had missing values, the data were preprocessed by excluding indexes with an absent rate greater than 10%, excluding the data of patients with missing values, and normalizing the remaining data.

Intraoperative data:Due to the varying sampling frequencies of the intraoperative data collection instruments, the sampling frequency of all data was adjusted to 1 Hz during preprocessing in this study. Original data collected by the anesthesia machine were sampled at 0.048 Hz, then data upsampling to 1 Hz was performed using data fill, i.e., the first sampled data was used to fill the gap between two samples, and finally, outlier handling was performed for the output indicators. Each indicator's values less than or equal to 0 were considered outliers. Indicators with identical values were deemed abnormal, whereas those with an abnormality rate exceeding 10% were eliminated. Regarding individual data outliers, the mean value of the indicator for that patient was used to fill in the blanks; the final indicators used are listed in Table S2 of the appendix. After preprocessing, data with lengths longer than 18,000 s were intercepted to 18,000 s, data with lengths shorter than 18,000 s were filled with zeros, and the data were then normalized.

③The cross-validation process used in the machine learning analysis:In this study, K-fold cross-validation was adopted, and k was set to 10, that is, the original data was randomly divided into 10 parts without repeated sampling. One of them is selected as the test set at a time, and the remaining nine are used as the training set to train the model. Repeat step 2 10 times to get a model after training on each training set. Test the model against the appropriate test set,

calculate and save the model's evaluation metrics. The average of the 10 groups of test results is calculated as an estimate of the accuracy of the model and as a performance index of the model under the current K-fold cross-validation.

This study actually uses machine learning as a tool to identify patients at high risk of OPCABG early and guide clinical decision-making during the perioperative period of OPCABG. It is more focused on obtaining clinically meaningful results through machine learning as a tool, but the specific process is not the main content of this paper. Nevertheless, we will supplement the specific training process of machine learning here, but we will not add too much in this article.

Changes in the text: Page 5/line 142-143; and Page6/line 159-166

Reviewer B

The authors aimed to establish a prediction model for OPCABG-AKI based on machine learning methods. They concluded that a GBDT-based model showed excellent performance in predicting OPCABG-AKI, and the fusion of pre- and intraoperative data can improve accuracy.

General Remarks:

I found the manuscript intriguing. Their endeavor to create a predictive model for surgery using machine learning methods is innovative and applicable across various surgical procedures. In cardiac surgery, particularly, postoperative kidney injury is critical as it directly impacts postoperative prognosis.

Itemized Remarks:

The ranking of feature importance in the GBDT model highlighted the use of acarbose, spironolactone, alfentanil, dezocine, levosimendan, clindamycin, etc. The authors suggest that these drugs may affect renal function after cardiac surgery.

Comment 1: Acarbose is used for diabetic patients. I wonder why diabetes mellitus did not emerge as a predictive factor.

Reply 1: The predictive factors obtained in this study were not based on clinical experience or artificial selection, but by calculating the accuracy and AUC of the prediction model, selecting the best model to predict OPCABG-AKI, and then using the DT model to calculate and rank the importance of features to screen out the main risk factors, and there was no diabetes in the results.

Changes in the text: None.

Comment 2: Alfentanil, dezocine, levosimendan, and clindamycin are not commonly used

preoperatively. I'm curious about the number of patients who received these drugs. There might be a correlation rather than a causal relationship.

Reply 2: The number of patients who received alfentanil is 221.The number of patients who received dezoucine is 111.The number of patients who received levosimendan is 137.The number of patients who received clindamycin is 571.Agree with the reviewer. Patients with these risk factors are not necessarily associated with acute kidney injury, but these risk factors are highly correlated with postoperative acute kidney injury. Therefore, anesthesiologists should also regard patients with predictive factors of OPCABG-AKI in this study as high-risk patients with postoperative acute kidney injury, and focus on their renal function. Actively give kidney protection related intervention measures.

Changes in the text: None.

Reviewer C

Comments: The manuscript is intriguing and timely. Data presentation is clear Methodology is sophisticated and outcomes definitions appropriate. The major limitations derive from the lack of intra procedural data (like number of anastomoses, type of grafts, conversion to on pump) that might be expressed as a surrogate by the physiological changes incorporated in the model. The discussion is well developed and reports pertaining references nevertheless the clinical bottom line of key findings might be improved.

Reply: Thanks for the advice of reviewer. This study does have some limitations, which I will address in the discussion to refine future research.

Changes in the text: Page 9/line 281-289.

Reviewer D

Here is my review of the manuscript "Machine learning-based prediction of off-pump coronary artery bypass grafting-associated acute kidney injury":

Strengths:

The study addresses an important clinical issue of predicting acute kidney injury (AKI) after off-pump CABG, which is a common and serious complication. Developing accurate prediction models could help identify high-risk patients for preventive interventions.

The authors compared the performance of 8 different machine learning algorithms and identified gradient-boosted decision trees (GBDT) as the best model, providing useful information on the comparative utility of these techniques for this prediction task.

Incorporating intraoperative time-series data extracted via a deep learning model substantially improved the predictive performance compared to using preoperative data alone. This supports the value of granular intraoperative data for risk prediction.
The authors used data from a reasonably large cohort of over 1000 patients, although the number of AKI events was limited (73 events).
Feature importance from the GBDT model provides some insights into potential risk factors for AKI after off-pump CABG.

Weaknesses:

Details on the study cohort are lacking. The authors should provide a flow diagram describing how many patients were screened, the reasons for exclusion, and the final sample size. Baseline characteristics of included patients should be reported.
The handling of missing data is not clearly described. More specifics are needed beyond excluding variables with >10% missing and imputing remaining missing values. The imputation method should be detailed.
Details on model development are inadequate. Issues like hyperparameter tuning, cross-validation approach, and how the final model was specified should be provided. Confidence intervals should be reported for performance metrics.
There is no discussion of limitations, like the single-center nature of the study, the small number of AKI events, and the lack of external validation, which limits generalizability.
The discussion section is cursory and should be expanded to put the results in the context of prior literature, discuss clinical implications, and outline future research directions.

The study addresses an important issue but has significant reporting deficits that make it difficult to evaluate fully. Major revisions are needed to add key methodological details, expand the discussion, and address limitations. Specific areas to address include:

Comment: Provide a flow diagram detailing cohort selection with reasons for exclusion. Report baseline characteristics of the final cohort.
Reply: The flow diagram detailing cohort selection with reasons for exclusion will be supplemented in this article. This study is not about simple component comparison, but focuses on machine learning model prediction. There is no problem of component comparison in this algorithm, and only two overall data sets, training set and test set, exist. The specific cases in the training and test sets were randomized for both the AKI and non-AKI groups. And the machine learning operation is constantly re-randomizing the combination of cases, each time the two groups of patients change, for example, case 1 in the first operation in the training set, maybe the second is randomized into the test set. Therefore, the case selection of machine learning is made by the machine randomly, each time is not repeated, the position of the cases

is not fixed, and the cases will not be selected by humans to enter the group. Therefore, the baseline characteristics cannot be provided.

Changes in the text: Page 6/line 181, and page 7/line 196.

Comment: Clarify all steps of data preprocessing, including imputation methods for missing data. Expand the description of model development, including hyperparameter selection, cross-validation methods, and final model specification.

Reply: ①Imputation methods for missing data:Due to the varying sampling frequencies of the intraoperative data collection instruments, the sampling frequency of all data was adjusted to 1 Hz during preprocessing in this study. Original data collected by the anesthesia machine were sampled at 0.048 Hz, then data upsampling to 1 Hz was performed using data fill, i.e., the first sampled data was used to fill the gap between two samples, and finally, outlier handling was performed for the output indicators. Each indicator's values less than or equal to 0 were considered outliers. Indicators with identical values were deemed abnormal, whereas those with an abnormality rate exceeding 10% were eliminated. Regarding individual data outliers, the mean value of the indicator for that patient was used to fill in the blanks; the final indicators used are listed in Table S2 of the appendix. After preprocessing, data with lengths longer than 18,000 s were intercepted to 18,000 s, data with lengths shorter than 18,000 s were filled with zeros, and the data were then normalized.

②Hyperparameter selection:The length of the preprocessed intraoperative dataset L was 18,000, and the number of feature data N was 28. During the intraoperative data segmentation stage, K was set to 180, P was set to 120, zero padding length was 90, final S was 150, and the 3-D tensor was obtained after segmentation combination.

Each deep learning model was pre-trained for 100 epochs on an 18,000-second sequence, and the training was terminated early if the best model was not found in the validation set for 30 consecutive epochs. The initial learning rate was 1e-3, and the decay rate was 0·97 every two epochs. The optimizer used was Adam30, and the pre-trainingloss function was the traditional Cross-Entropy Loss.Accuracy and area under the curve (AUC)31 were calculated for each model (both machine learning models and deep learning models), with AUC serving as the primary evaluation metric for determining the model's performance. In order to address the issue of an unbalanced class, the loss function's weight was also added to the models.

③Cross-validation methods:In this study, K-fold cross-validation was adopted, and k was set to 10, that is, the original data was randomly divided into 10 parts without repeated sampling. One of them is selected as the test set at a time, and the remaining nine are used as the training set to train the model. Repeat step 2 10 times to get a model after training on each training set. Test the model against the appropriate test set, calculate and save the model's evaluation metrics. The average of the 10 groups of test results is calculated as an estimate of the accuracy of the model and as a performance index of the model under the current K-fold cross-validation

④Final model specification:Accuracy and Area under the Curve (AUC) 31 Each model, including machine learning models and deep learning models, is calculated using the AUC as the primary evaluation metric to determine the model's performance. In order to solve the problem of unbalanced class, the weight of loss function is added to the model.
This study actually uses machine learning as a tool to identify patients at high risk of OPCABG early and guide clinical decision-making during the perioperative period of OPCABG. It is more focused on obtaining clinically meaningful results through machine learning as a tool, but the specific process is not the main content of this paper. Nevertheless, we will supplement the specific training process of machine learning here, but we will not add too much in this article.
Changes in the text: Page 5/142-143, and Page6,159-166

Comment: Report 95% confidence intervals for model performance metrics and consider formal comparisons between models.
Reply: The 95% confidence intervals are in statistical analysis, machine learning doesn't think about them.
Changes in the text: None.

Comment: Expand the discussion to address how your findings compare to prior studies, implications for clinical practice, limitations of your study, and future research needs.
Reply: Thanks for the advice of reviewer. I will add how your findings compare to prior studies, implications for clinical practice, limitations of your study, and future research needs in this article.
Changes in the text: Page 9/line 262-270.